LAGNet: A Location-Aware Guidance Network for Weak and Strip Defect Detection

Lisha Cui, Helong Jiao, Tengyue Liu, Chunyan Niu*, Ming Ma, Xiaoheng Jiang, Mingliang Xu
School of Computer Science and Artificial Intelligence, Zhengzhou University
Zhengzhou 450001, Henan, China
ielscui@zzu.edu.cn, iehljiao@163.com, ietyliu@gs.zzu.edu.cn, yanzi198929@163.com,
maming@gs.zzu.edu.cn, iexhjiang@zzu.edu.cn, iexumingliang@zzu.edu.cn

Abstract

Automatically surface defect detection plays a crucial role during the industrial production process. Unfortunately, some special defects, such as weak and strip-like defects, are relatively difficult to classify and localize accurately. In this paper, we propose a novel Location-Aware Guidance Network for weak and strip defect detection, termed as LAGNet. To enhance the feature representation of weak defects, we introduce the Location Activation Map (LAM) by visualizing the confidence score map that indicates the probability of the existence of an object in each region. The LAM and RGB images are fed into the network in a parallel manner for feature extraction, and then we fuse these two branches via a Location Guidance Block (LGB) that inherently encodes comprehensive and complementary information for detection. Additionally, a Strip Convolution Enhancement Module (SCEM) is presented using the depthwise strip convolutions with long but narrow kernels and attention mechanism and plugged into the detection neck to model the long-range dependencies along both horizontal and vertical spatial directions, thus improving the detection performance of anisotropic defects with banded structures. Notably, LAGNet achieves the top-ranking results on two popular steel benchmarks and significantly outperforms the baseline network YOLOv5: 85.5% mAP (*vs.* 76.2%) on NEU-DET and 76.6% mAP (*vs.* 66.0%) on GC10-DET.

Keywords: Defect detection, Weak defect, Location activation maps, Strip convolution, Attention mechanism.

1. Introduction

With the continuous evolution of industrial production, steel, as a key structural material, is widely utilized in various types of infrastructure and engineering projects. However, diverse surface defects accidentally occur during the manufacturing and processing of steel, which seriously jeopardizes the quality and service life. Therefore, rapid and accurate detection of these defects has become a key assurance of product quality and safety.

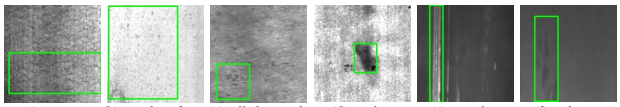


Figure 1. Examples of steel surface defects in NEU-DET dataset.

Traditional manual defect inspection is not only time-consuming and labor-intensive but also susceptible to subjective factors that lead to high false and missed detection rates. Classic automatic defect detectors [1] [2] [3] [4] rely on hand-crafted features, which fail to deal with the diversity of defects and have high requirements for image quality as well. Recent advances based on deep learning techniques, especially Convolutional Neural Networks (CNN), have achieved remarkable performances and occupied a dominant position in the field of defect detection [5] [6][7]. Such detectors have been proven to have stronger generalization ability and adaptability.

Although defect detectors based on CNN have shown significant advantages, the performance is still less than satisfactory due to the complexity of defect images. Compared with general object detection, defect detection usually has the following challenges: (1) The subtle texture difference between defective and defect-free areas in RGB images. Due to the randomness in the production process and objective factors in the data acquisition process, such as lighting conditions, certain defects exhibit weak features, low contrast, and indistinct boundaries, as shown in Figure 1 (a), (b), and (c). This makes it more challenging to accurately classify and localize defects, resulting in a higher rate of false positives and false negatives. (2) The extremely large aspect ratios of the defects with banded structures. Unlike the objects in Pascal VOC [8], some of the defects are in long and thin shapes, such as *scratches* and *inclusion* shown in Figure 1 (e) and (f). It is difficult for the detectors to deal

with the defects with arbitrary aspect ratios.

To cope with the first issue, it makes very logical sense for the network to encode and enhance the location information of weak defects in RGB images. Consequently, we purposely introduce the location activation maps by visualizing the importance of different regions for the localization task in the form of color-coded heat maps. The brighter color on the maps indicates the stronger response and greater contribution of each point to the localization task. The RGB image and its corresponding activation map are fed into the network in a parallel fashion to encode respective features. Besides, a location guidance block is proposed to integrate the outputs of the two branches in the middle and late stages and guide the network to capture comprehensive and complementary features, especially strengthening the location information and improving the recall of weak defects in RGB images.

For the strip objects in defective images, the standard square kernel fails to model their long-range spatial dependencies, and may also incorporate irrelevant contaminating information from the background. Based on this observation, we propose a strip convolution enhancement module based on strip convolutions and attention mechanisms to improve the modeling ability of CNN for long-distance spatial information. The strip convolutions, such as 1×5 or 5×1, enable the network to capture long-range context along the horizontal or vertical directions. Moreover, following the strip convolutions, we introduce channel and spatial attention to further strengthen the feature representation of weak defects.

The strip convolution enhancement module performs finer and more comprehensive feature extraction for elongated defective regions in both horizontal and vertical directions, thus improving defect detection performance, especially the weak and slender defects. Also notably, the module is lightweight and flexible to be plugged into any network.

In summary, the main contributions of this work can be summarized as follows:

• We innovatively introduce the Location Activation Map (LAM) as the auxiliary data to compensate for the poor location information in RGB images. Meanwhile, the Location Guidance Block (LGB) is tailored exclusively for effectively integrating the parallel RGB and LAM branches, guiding the network to encode more comprehensive and discriminative features for weak defect detection.

• We propose a computation-friendly Strip Convolution Enhancement Module (SCEM) based on the long but narrow kernel and attention mechanisms, modeling the long-range dependencies along horizontal and vertical directions while focusing on local details of defects.

• We present a novel Location-Aware Guidance Network for weak and strip defect detection, abbreviated as LAGNet. LAGNet consistently produces competitive results on the widely used defect benchmarks NEU-DET (85.5% mAP) and GC10-DET (76.6% mAP). Most notably, the small version of LAGNet achieves 84.9% mAP and reaches 73.6 FPS on NEU-DET with only 8.9M parameters.

2. Related work

2.1. General object detection

Deep learning has achieved great success in the field of computer vision, especially with the emergence of CNNs which have greatly improved the performance of general object detection. Currently, object detectors based on deep learning can be roughly divided into two- and one-stage object detection.

The R-CNN [9] [10] [11] family is the classical two-stage object detector. R-CNN [9] applies a region proposal network (RPN) to generate candidate object regions, then extracts the features of each region by CNN, and finally adopts a classifier and regressor to classify and localize the object. Instead of generating candidate regions first, one-stage object detectors accomplish object localization and classification directly in a single neural network, as represented by the SSD [12] and YOLO [13] [14] [15] [16] [17] [18] series. SSD adopts multi-scale feature maps to predict objects with different scales, and directly predict the confidence scores and box offsets through convolutions. The YOLO series divides the image into fixed-size grid cells and predicts bounding boxes and category probabilities in each cell. YOLOv4 [16] is an efficient object detection model that combines multiple optimization techniques to improve detection accuracy while maintaining real-time speed. YOLOv5 [17] further improves the training strategies and optimized loss functions, making it an ideal choice for real-time applications. YOLOv7 [18] introduces model reparameterization, label assignment strategies, extended efficient layer aggregation networks (ELAN), and auxiliary head training techniques.

Although these classic general object detectors have achieved excellent results on routine tasks, they have difficulty in accurately detecting defects with faint features, low contrast, and arbitrary aspect ratios.

2.2. Defect detection

The successful application of CNN in the field of image classification and object detection provides a brand new direction for defect detection. Therefore, defect detection algorithms based on CNN have become a hot research topic for industrial quality inspection at present.

Zhao *et al*. [19] innovatively design a double feature pyramid network (DFPN) based on Res2Net [20] to increase the semantic information. Liu *et al*. [21] propose a parallel architecture of dilated convolution (PADC) with

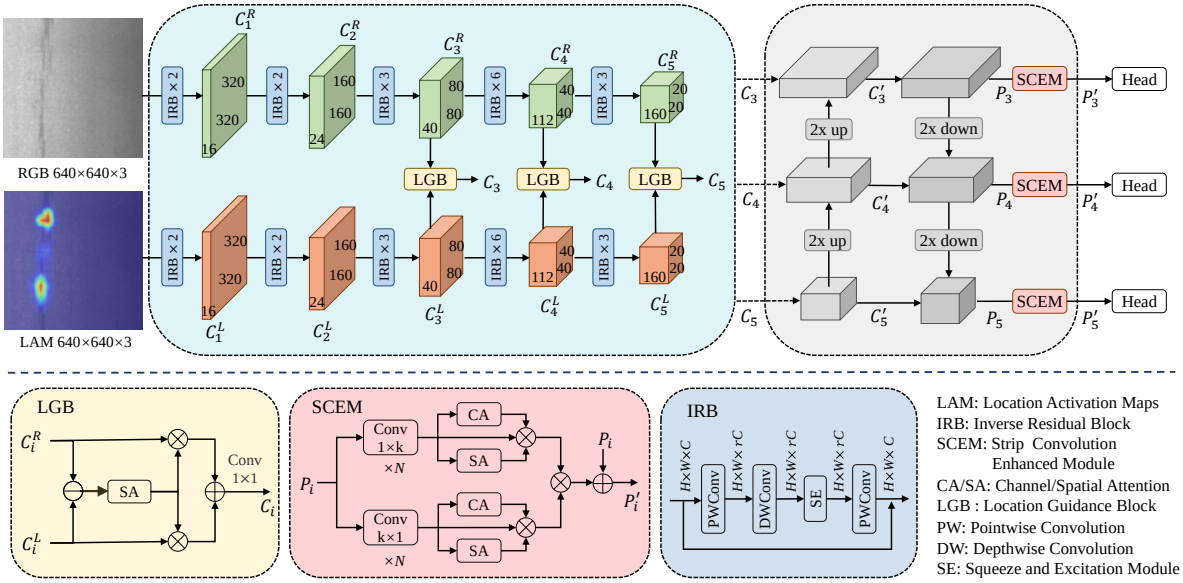


Figure 2. The overall structure of proposed LAGNet, which consists of the dual-branch backbone, the feature pyramid neck, and a multi-scale detection head. The LAM could provide RGB auxiliary rich location information and guide the network to focus on the discriminative defective areas through LGB. The SCEM deployed in the detection neck adopts the strip convolutions and attention mechanism to model the long-range spatial dependencies of strip defects.

different dilation rates to capture multi-scale contextual information and a feature enhancement and selection module (FESM) to enhance single-scale features. On the basis of CenterNet [22], Tian *et al*. [23] propose an anchor-free framework, DCC-CenterNet, for steel surface defect detection by introducing the dilated feature enhancement model (DFEM) and the centerness function center-weight (CW). In addition, to cope with the problem of visual defect detection in complex images, Yu *et al*. [24] propose a progressively refined redistribution pyramid network.

Although the aforementioned algorithms have improved the overall accuracy of defect detection, their performance is still unsatisfactory when facing weak defects with low contrast. To overcome this challenge, we employ the location activation maps as a localization guide to motivate the network to focus on the weak defective areas.

2.3. Strip convolution

Beyond the regular convolutional kernel with a square shape of $k \times k$, strip convolutions apply long but narrow kernels, such as $1 \times k$ or $k \times 1$, to efficiently model long-range dependencies. ACNet [25] introduces strip convolution blocks, replacing the common 3×3, 5×5, and 7×7 square kernels, to effectively support the extraction of certain asymmetric image features. Strip Pooling [26] captures long-range relationships in isolated regions by performing pooling operations along the horizontal or vertical dimension.

In InceptionNeXt [27], the large-kernel depth convolution is decomposed into four parallel branches, including small square kernels, two strip convolution kernels, and a unit map, to achieve more flexible feature extraction. SegNeXt [28] employs two depth strip convolutions to approximate standard depth convolutions with large kernels. Depth strip convolution is lightweight and only requires a pair of $k \times 1$ and $1 \times k$ convolutions to imitate a standard convolution with a square kernel size of $k \times k$.

Based on the above findings, in this paper, we present a strip convolution enhancement module to capture long-range context and focus on local details, thus improving the detection performance of strip defects.

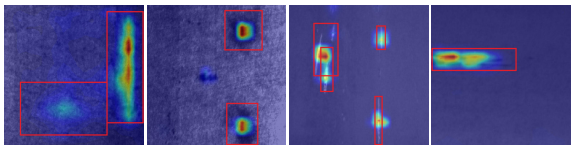
3. Proposed method

In this section, we first introduce the overall pipeline of proposed LAGNet and the generation of LAM. Subsequently, we describe the backbone network with two parallel branches of RGB and LAM in detail, as well as the fusion module LGB. Finally, we elaborate on the structure of SCEM.

3.1. Overall structure

Figure 2 illustrates the overall framework of the proposed LAGNet, which mainly consists of three parts: a two-pathway backbone, a feature pyramid neck, and a multi-scale detection head.

The backbone network is composed of two parallel branches, each with RGB images and LAM as input re-



(a) crazing (b) patches (c) scratches (d) inclusion

Figure 3. The LAM generated by the well-trained YOLOv5.

spectively. Considering the edge end employment in embedded devices, we choose the lightweight MobileNetV3 [29] to extract features from both RGB images and LAM simultaneously. The corresponding output feature maps of each branch in the middle and late stages undergo the LGB and integrate together to learn complementary signals, producing location-enhanced feature maps C_3, C_4, and C_5. Subsequently, these feature maps are fed into a top-down and bottom-up bidirectional pathway to build high-level and strong semantic features P_3, P_4, and P_5. Notice that the proposed SCEM is placed after the multiple hierarchical features to efficiently model long-range dependencies and produce P_3', P_4', and P_5'. Finally, the predictions are made on these multi-scale feature maps to deal with large variants of defect sizes.

3.2. The generation of LAM

The faint defects present low contrast and small texture differences from the background in RGB images, which makes it a huge challenge for the network to recognize and localize the weak defects accurately. Therefore, we introduce the location activation maps.

Specifically, we first fine-tune a popular detector (such as YOLOv5) on the defect benchmarks and achieve the well-trained models. For each input image, the well-trained model predicts the box coordinates, classes, and confidence scores of defects. The confidence score map (CSM) indicates the probability of the existence of an object in each region, where the higher the value, the greater the possibility that an object exists. We first normalize the CSM output by three YOLOv5 detection heads to [0, 1] using the sigmoid activation function, and then multiply them by 255. Then a location information map is obtained by selecting the maximum value among the processed three CSMs with different scales. Finally, the color-coded LAM is produced by fusing the location information map with its corresponding image in a ratio of 1:1, as shown in Figure 3. Mathematically, the generative process can be formulated as follows:

$$LAM = \alpha(RGB) \oplus \beta(255 \otimes sigmoid(CSM)) \quad (1)$$

where RGB and CSM represent the input RGB image and corresponding confidence score map, respectively. α and β refer to the proportional coefficients when fusing RGB and CSM, and both are set to 0.5 after experimental trials. $\oplus$ and $\otimes$ refer to element-wise summation and multiplication,

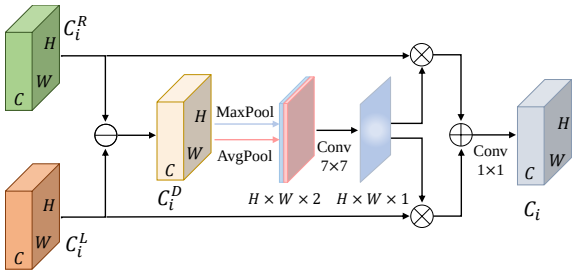


Figure 4. Illustration of the proposed LGB.

respectively. As illustrated in Figure 3, the higher brightness on LAM indicates that this region contributes more to the localization task. Therefore, the LAM could highlight the defect area and compensate for the location information of weak defects in RGB images.

3.3. The dual-branch backbone

RGB images are rich in texture, color, and shape information, whereas susceptible to interference from background noise, light, *etc*. Conversely, the LAM could highlight the defective areas and enhance the contrast with the background. Fully utilizing the complementary information of both RGB images and LAM is beneficial for improving defect detection performance, especially the weak defects.

Based on the above considerations, we propose the LAGNet for weak defect detection with a dual-branch backbone, *i.e.* RGB and LAM branches. As illustrated in Figure 2, the RGB images and corresponding LAM are fed forward into the network to encode features of input data separately. Then we specially design the LGB that fuses the output features of both branches in the middle and late stages to enhance the location details and construct more comprehensive feature maps. Figure 4 depicts the structure of the proposed LGB.

Specifically, imagine C_i^R, $C_i^L \in \mathbb{R}^{H \times W \times C}$ are the feature maps output by RGB and LAM branches in stage i, where H, W, and C denote the height, width, and number of channels, respectively. First, let C_i^R and C_i^L perform subtraction to enhance the visible features and produce the differential-modality feature maps C_i^D $(i = 3, 4, 5)$. Subsequently, the spatial attention module is applied to further enhance the defect feature representation, and the attention map is then multiplied by C_i^R and C_i^L, respectively. Finally, the dual branches are integrated together by element-wise summation and pass through a 1×1 convolution layer to refine the fusion features, producing $C_i \in \mathbb{R}^{H \times W \times C}$. The process can be described as follows:

$$C_i^D = C_i^R \ominus C_i^L, i = 3, 4, 5 \quad (2)$$

$$C_i = Conv_{1\times1}((C_i^R \otimes SA(C_i^D)) \oplus (C_i^L \otimes SA(C_i^D))) \quad (3)$$

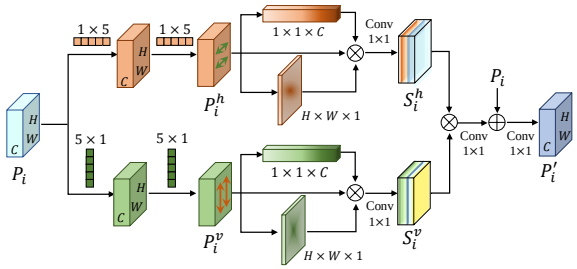


Figure 5. Illustration of SCEM. The upper and lower branches aim to model the spatial relations along the horizontal and vertical directions, respectively.

where $\ominus$ refers to element-wise subtraction. SA stands for the spatial attention module.

The LGB could guide the network to pay more attention to defective areas, and thus produce more informative and representative feature pyramids C_3, C_4, and C_5. Considering the dual-branch backbone in LAGNet inevitably introduces extra parameters and computational complexity, we choose the efficient and lightweight MobileNetV3 as the backbone for feature extraction. MobileNetV3 is based on Inverse Residual Block (IRB) shown in Figure 2, whose feature space remains constant at the input and output while internally expanding to a higher dimension. Additionally, IRB replaces the traditional convolution layers with depthwise separable convolutions to efficiently trade off between latency and accuracy.

3.4. Strip Convolution Enhancement Module

The box regression of strip defects, such as *scratches* and *inclusion*, is a challenging task compared with nearly square-shaped defects. The standard convolution with a regular shape of $k \times k$ (such as commonly used 3×3) fails to capture long-range contextual information. A larger kernel (such as 7×7) is able to encode more comprehensive information, but may introduce contaminating information from irrelevant background, as well as additional parameters.

Regarding this problem, we devise the SCEM based on the depth-separable strip convolution with long but narrow kernels and plug it into the detection neck to model the long-range dependencies, as demonstrated in Figure 5. Specifically, the feature hierarchies C_3, C_4, and C_5 generated by LGB first undergo a top-down and bottom-up pipeline to encode high-level semantic features, producing P_3, P_4, and P_5, respectively.

$$C_5^{'} = Conv_{1\times1}(C_5), C_i^{'} = Up(C_{i+1}^{'}) \oplus C_i, i = 3, 4 \quad (4)$$

$$P_3 = Conv_{1\times1}(C_3^{'}), P_i = Down(P_{i-1}) \oplus C_i^{'}, i = 4, 5 \quad (5)$$

P_3, P_4, and P_5 are then fed into SCEM to capture long-range relations of local regions. More concretely, taking $P_i \in \mathbb{R}^{H\times W\times C}$ $(i = 3, 4, 5)$ as input, SCEM consists of two parallel pathways as shown in Figure 5. The upper branch first deploys n stacked depthwise strip convolutions with the size of $1 \times k$ to encode global context along the horizontal spatial dimension. Similarly, the lower branch is to capture global vertical information using n consecutive $k \times 1$ convolutions. By conducting ablation experiments in Section 4.5, we set $n = 2$ and $k = 5$ in this study. The intermediate output feature maps of the upper and lower branches in this step are P_i^h and P_i^v respectively, which can be described as:

$$P_i^h = DWConv_{1\times k}(DWConv_{1\times k}(P_i)), i = 3, 4, 5 \quad (6)$$

$$P_i^v = DWConv_{k\times 1}(DWConv_{k\times 1}(P_i)), i = 3, 4, 5 \quad (7)$$

where $DWConv$ denotes the depthwise strip convolution and P_i is the input feature map.

To further enhance the feature representation for weak defects, P_i^h and P_i^v go through the attention mechanism in both the channel and spatial dimensions. The channel tensor $1 \times 1 \times C$ predicts the importance of each channel and pays attention to what is in the image, while the spatial map $H \times W \times 1$ indicates the significance of each location and focuses on where is the defect. The channel and spatial attention module are arranged in a parallel fashion, and then multiplied by P_i^h and P_i^v, respectively. As usual, a pointwise convolution $1 \times 1 \times C$ is followed to compute the linear combination of the output from preceding layers, fully mixing inter-channel information and producing S_i^h and S_i^v which can be formulated as:

$$S_i^h = PWConv_{1\times1}(P_i^h \otimes CA(P_i^h) \otimes SA(P_i^h)) \quad (8)$$

$$S_i^v = PWConv_{1\times1}(P_i^v \otimes CA(P_i^v) \otimes SA(P_i^v)) \quad (9)$$

where $PWConv$ represents the pointwise convolution. CA and SA are the channel and spatial attention, respectively.

Finally, we perform an element-wise multiplication of S_i^h and S_i^v, and then introduce a residual connection with P_i by element-wise summation. Both of these fusion operations are followed by one $1 \times 1 \times C$ convolution to refine the features, generating the more comprehensive P_i'. Mathematically, P_i' can be written as:

$$P_i' = Conv_{1\times1}(P_i \oplus Conv_{1\times1}(S_i^h \otimes S_i^v)), i = 3, 4, 5 \quad (10)$$

SCEM considers the long but narrow spatial dependencies along both horizontal and vertical directions over the whole scene, thus improving the capability of modeling long-range information for strip defects. Simultaneously, the depth-separable convolutions considerably reduce the computational complexity and module size, making SCEM lightweight and easily plugged into any architecture.

The output feature maps of SCEM, P_i' $(i = 3, 4, 5)$, inherently encode long-range spatial dependencies and discriminative features, which are then utilized to predict multi-scale defects.

3.5. Training objective

The training objective of LAGNet is the weighted sum of the box regression loss, confidence loss, and classification loss. The box loss (L_{box}) is the CIoU loss:

$$L_{box} = L_{CIoU} = 1 - IoU + \frac{\rho^2(b, b^{gt})}{c^2} + \alpha v \tag{11}$$

$$IoU = \frac{|b \cap b^{gt}|}{|b \cup b^{gt}|} \tag{12}$$

$$v = \frac{4}{\pi^2}\left(\arctan\frac{w}{h} - \arctan\frac{w^{gt}}{h^{gt}}\right)^2 \tag{13}$$

where b and b^{gt} represent the predicted and ground-truth boxes, respectively. ρ refers to the distance between the center points of the two bounding boxes, and c is the diagonal length of the smallest outer rectangle of the two boxes. α is the weight coefficient and is usually set to 0.5. w and h denote the width and height of the predicted box, while w^{gt} and h^{gt} are the width and height of the ground-truth box.

The confidence loss (L_{obj}) and classification loss (L_{cls}) are Binary Cross-Entropy (BCE) Loss. The total loss function L is the weighted sum of these three parts:

$$L = \lambda_{\text{box}} L_{\text{box}} + \lambda_{\text{obj}} L_{\text{obj}} + \lambda_{\text{cls}} L_{\text{cls}} \tag{14}$$

where λ_{box}, λ_{obj}, and λ_{cls} are the weight coefficients and set to 0.05, 1, and 0.5, respectively.

4. Experiments

In this section, we first elaborate the implementation details and evaluation metrics of the experiments and introduce the defect datasets employed in this paper. Subsequently, we present comprehensive comparisons of results regarding LAGNet and the current SOTA methods. Finally, we analyze and validate the effectiveness of each component in LAGNet through extensive ablation studies.

4.1. Implementation details

The proposed LAGNet and other comparison algorithms are carried on the PyTorch framework. We implement all the experiments on a computing platform equipped with an NVIDIA GeForce RTX4090 GPU. The stochastic gradient descent (SGD) algorithm is utilized to optimize the model parameters, where the momentum and weight decay are set to 0.937 and 0.0005, respectively. The initial learning rate is set to 0.01, and a warm-up strategy and cosine annealing learning rate strategy are introduced with a total training epoch of 500. The batch size is uniformly set to 16. During the training process, we apply data augmentation techniques such as Mosaic and Mixup for both RGB images and LAM to improve the generalization ability of the models. Moreover, LAGNet is based on the MobileNetV3-Large as the backbone for feature extraction, while a small version is created by adopting MobileNetV3-Small to meet industrial demands, namely LAGNet-s.

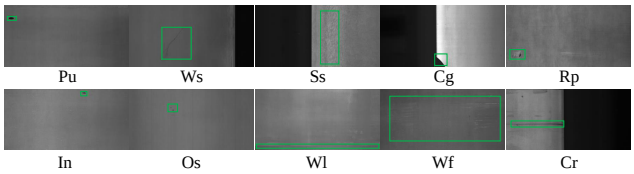


Figure 6. Examples of defects in the GC10-DET dataset.

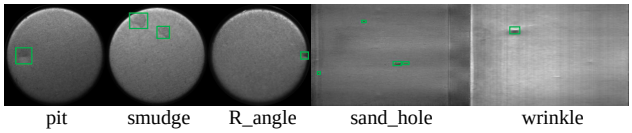


Figure 7. Examples of defects in the GB-DET dataset.

4.2. Datasets

We validate the proposed LAGNet and LAGNet-s on two publicly available defect detection datasets (NEU-DET [30] and GC10-DET [31]) and one self-constructed dataset (GB-DET). **NEU-DET** (Figure 1) contains 6 classes of common steel surface defects: *patches*, *crazing*, *rolled-in_scale*, *pitted_surface*, *inclusion*, and *scratches*. Each category comprises 300 images, making a grand total of 1800 defect images, each with a size of 200 × 200. The whole dataset is randomly partitioned into the training and test set with a ratio of 7:3 in [30], namely 1260 images for training and 540 images for testing. **GC10-DET** shown in Figure 6 contains 10 types of defects: *punching* (*Pu*), *welding line* (*Wl*), *crescent gap* (*Cg*), *water spot* (*Ws*), *oil spot* (*Os*), *silk spot* (*Ss*), *inclusion* (*In*), *rolled pit* (*Rp*), *crease* (*Cr*), and *waist folding* (*Wf*). There are 2294 images in total, each with the size of 2048 × 1000 pixels. Following other algorithms, we randomly divide the whole dataset into the training and test set with a ratio of 9:1 [32], namely 2064 images for training and 230 images for testing. The images in both datasets are scaled to 640 × 640 as input during training and testing. Among these 10 types of defects, *Ss*, *Wl*, and *Cr* are typically strip-like defects. **GB-DET** is a self-constructed dataset for surface defect detection of renewable energy batteries, which contains 5 types of defects: *pit*, *smudge*, *R_angle*, *sand_hole*, and *wrinkle*. As shown in Figure 7, the images containing circular bottom shells have a resolution of 320 × 320, whereas the images containing rectangular side shells have a resolution of 800 × 512. There are 6368 images in total, and we randomly divide them into the training and test set with a ratio of 7:3, namely 4458 images for training and 1910 images for testing.

Table 1. Detection results of state-of-the-art methods on NEU-DET.

Method	Backbone	FPS	Params	mAP(%)	crazing	inclusion	rolled-in_scale	scratches	patches	pitted_surface
Object detectors										
Faster R-CNN [11]	ResNet50	22.3	28.4M	77.9	52.5	76.5	74.4	90.3	89.0	84.7
SSD512 [12]	VGG16	64.9	27.0M	72.1	39.9	79.6	61.9	84.4	86.7	79.8
YOLOv5 [17]	CSPDarkNet	109.3	46.1M	76.2	38.4	82.0	64.5	94.0	**94.8**	83.3
YOLOv7 [18]	-	62.5	36.5M	75.9	42.0	83.9	63.0	89.1	94.3	83.0
YOLOX [33]	CSPDarkNet	53.7	54.2M	76.8	47.6	81.7	60.9	94.4	92.9	83.3
YOLOv10 [34]	CSPDarkNet	161.1	24.4 M	75.9	40.6	87.1	67.4	91.5	84.2	84.5
Defect detectors										
DIN [35]	-	-	-	80.5	61.4	85.6	64.6	88.3	93.0	90.3
RDN [36]	ResNet18	-	-	80.0	53.7	84.9	64.4	95.9	93.8	87.0
DDN [30]	ResNet50	-	-	82.3	62.4	84.7	76.3	90.1	90.7	89.7
EFD-YOLOv4 [37]	CSPDarkNet	-	-	79.9	45.7	85.4	72.7	93.6	97.0	85.0
Zhang's [38]	CSPDarkNet	-	-	78.2	40.6	**90.3**	60.7	**96.1**	94.4	86.9
Multi-branch detectors										
CFT [39]	CSPDarkNet	23.8	206.2M	83.1	60.5	83.8	**82.5**	93.1	92.7	86.2
ICAFusion [40]	CSPDarkNet	24.5	120.3M	83.0	62.3	81.8	79.7	94.3	91.9	88.2
SuperYOLO [41]	CSPDarkNet	99.1	9.9M	78.9	54.9	80.1	72.1	86.9	91.4	87.2
Ghost [42]	CSPDarkNet	54.1	7.7M	72.0	46.5	72.2	72.3	71.7	89.3	80.9
LAGNet (Ours)	MobileNetV3	45.1	31.9M	**85.5**	**64.2**	86.4	82.4	**96.1**	94.4	89.2
LAGNet-s (Ours)	MobileNetV3	73.6	8.9M	84.9	63.6	85.3	80.5	95.5	94.4	**90.9**

4.3. Evaluation metrics

We utilize the average precision (AP), mean average precision (mAP), number of model parameters (Params), and frame per second (FPS) to evaluate the performance of different algorithms on NEU-DET and GC10-DET. The AP denotes the average accuracy of the model about a certain class, which is measured by the area under the precision-recall curve. The mAP represents the mean accuracy of the model for all classes, which is calculated by the average value of all APs. AP and mAP are defined as follows:

$$P = \frac{TP}{TP + FP}, R = \frac{TP}{TP + FN} \tag{15}$$

$$AP = \int_0^1 P(R)dR \tag{16}$$

$$mAP = \frac{1}{N} \sum_{i=1}^{n} AP_i \tag{17}$$

where P and R represent the precision and recall. TP, FP, and FN refer to true positives, false positives, and false negatives, respectively. The Params and FPS are to evaluate the model size and inference speed, respectively. The threshold of intersection over union (IoU) is set to 0.5. Besides, we also draw the Precision-Recall (PR) curves to evaluate the classification performance across different threshold settings. These metrics help to fully and comprehensively evaluate the performance of different methods.

4.4. Comparisons with the SOTA detectors

To demonstrate the effectiveness of the proposed method, we compare it with state-of-the-art detectors, including general object detectors, defect detectors, and multi-branch object detectors on both NEU-DET and GC10-DET datasets.

4.4.1 Results on NEU-DET

Quantitative results. Table 1 shows the extensive results of the proposed LAGNet and other state-of-the-art detectors on NEU-DET. It is observed that LAGNet achieves the highest 85.5% mAP, with an increment of 9.3% over the baseline YOLOv5 (76.2%). Compared with other RGB-only detectors, such as classic YOLOX (76.8%) and defect detector DDN (82.3%), LAGNet still has significant performance improvements. Although the multi-branch RGB- and LAM-based detectors, such as CFT (83.1%) and ICAFusion (83.0%), are of relatively higher accuracy, they have too many parameters to be deployed and applied. Fortunately, the proposed LAGNet can still exceed such detectors over 2 points with fewer parameters.

Notably, LAGNet-s substantially reduces the model size from 31.9M to 8.9M compared with LAGNet, which can be deployed on a computationally limited platform. Although the mAP of LAGNet-s (84.9%) has decreased slightly in comparison with LAGNet (85.5%), it can still outperform the lightweight SuperYOLO (78.9%) and Ghost (72.0%) by large margins with approximate model size. Additionally, LAGNet and LAGNet-s consistently achieve better results in terms of the AP. Especially for weak defect *crazing*, the AP of LAGNet is significantly improved to 64.2%, which lies in the enhanced location information from the LAM branch. Similarly, for *scratches* that have long and thin shapes, LAGNet encodes the long-range spatial dependencies by SCEM and further elevates the AP to 96.1%. The experimental results demonstrate the effectiveness of LAGNet for defect detection, especially the weak and strip

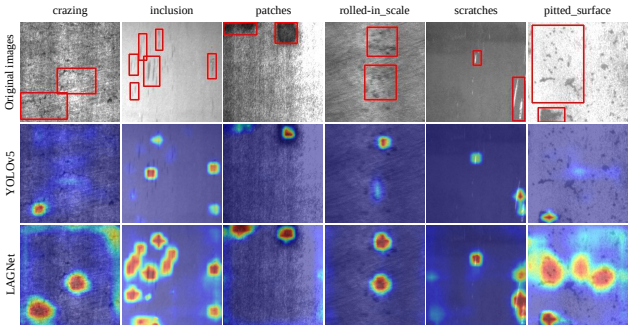


Figure 8. Comparison of activation maps generated by YOLOv5 and LAGNet on NEU-DET.

defects.

We also evaluate the inference speed and time efficiency of different algorithms on NEU-DET, as shown in the third column of Table 1. LAGNet introduces an additional LAM branch and extra parameters, so it inevitably impairs the model efficiency compared with the baseline network (109.3 FPS). Even so, LAGNet and LAGNet-s could still achieve 45.1 and 73.6 FPS, respectively. Especially, LAGNet-s could achieve the best trade-off among the detection accuracy, model size, and time efficiency in comparison with prevailing detectors.

Qualitative results. To visually demonstrate the localization ability of the models, we draw the activation maps using the well-trained YOLOv5 and LAGNet on NEU-DET, as shown in Figure 8. The topmost row is the defect images of 6 classes with ground-truth boxes. By contrast, it is observed that the proposed LAGNet could highlight the defective areas more completely and comprehensively compared with YOLOv5. We believe that the LGB enables the network to effectively focus on and enhance the representations of weak defects by supplementing location information, thus considerably improving the localization performance of the model. Figure 9 presents the comparison of the PR curves for different detectors. It is observed that the proposed LAGNet achieves a better trade-off between precision and recall, demonstrating superior classification performance compared with other algorithms.

Furthermore, we display the visualization detection results of LAGNet compared with YOLOv5 (general detector), CFT (multi-branch detector), and Zhang's (defect detector) on NEU-DET in Figure 10. The prediction boxes of different classes are distinguished by different colors. It can be visualized that LAGNet can accurately locate the weak defects in the low contrast background, such as *crazing* in the first column. Likewise, our method can also predict the strip defects in the fifth column more precisely than other detectors, demonstrating the effectiveness of SCEM in capturing long-range spatial dependencies.

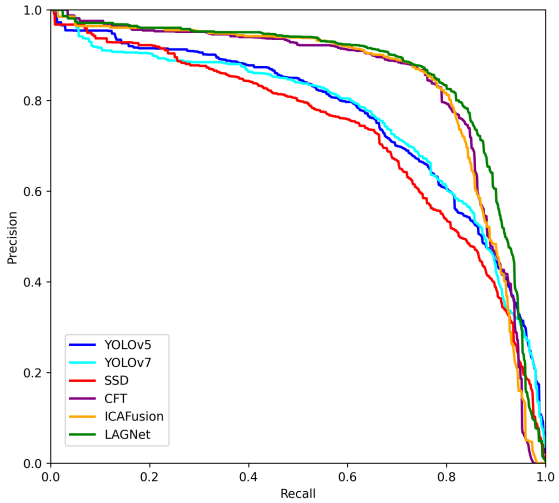


Figure 9. Comparison of PR curves for different models on NEU-DET.

4.4.2 Results on GC10-DET

To further verify the effectiveness and robustness of LAGNet, we also conduct experiments on GC10-DET, and the comparison results are illustrated in Table 2. LAGNet and LAGNet-s achieve 76.6% and 75.4% mAP, boosting the baseline YOLOv5 (66.0%) by 10.6 and 9.4 points, respectively. For general object detectors, YOLOv7 yields a desirable result with a mAP of 72.4%, which is the best result among YOLO families. For defect detectors, Zhang's obtains the excellent mAP of 71.9% using CSPDarkNet as the backbone. For multi-branch detectors, although CFT gets the impressive 74.4% mAP, it is under the guidance of the Transformer [43] scheme with a large model size (206.2M). Fortunately, LAGNet and LAGNet-s could outperform these different types of detection algorithms with acceptable parameter quantity.

It is noticeable that the APs of other methods vary greatly within a wide range, which may result from the imbalance of data distribution in GC10-DET. Despite that, LAGNet has yielded relatively stable APs for all defect classes. Additionally, Figure 11 depicts the visualization results of LAGNet on GC10-DET, where the weak defect *Rp* and slender defect *Cr* could be successfully predicted. Notably, LAGNet produces more accurate and compact bounding boxes than YOLOv5, which indicates that our model is more sensitive to the location of defects due to the introduction of LAM.

4.4.3 Results on GB-DET

We further evaluate LAGNet on our own dataset GB-DET with more defective images. The results are shown in Table 3. LAGNet achieves 93.2% mAP, which is superior to

Table 2. Detection results of state-of-the-art methods on GC10-DET.

Method	Backbone	Params	mAP(%)	Pu	Wl	Cg	Ws	Os	Ss	In	Rp	Cr	Wf
Object detectors													
Faster R-CNN [11]	ResNet50	28.4M	59.6	91.0	38.9	87.9	**79.7**	59.9	60.1	31.3	40.5	49.3	57.5
SSD512 [12]	VGG16	27.0M	60.6	95.7	91.7	**96.7**	66.6	60.8	45.7	16.1	22.1	26.1	**84.6**
YOLOv5 [17]	CSPDarkNet	46.1M	66.0	82.9	75.4	93.1	75.9	66.5	73.2	45.4	26.0	44.2	77.5
YOLOv7 [18]	-	36.5M	72.4	91.0	97.1	93.0	79.0	70.0	74.0	46.0	31.0	67.0	74.0
YOLOX [33]	CSPDarkNet	54.2M	71.0	84.2	96.1	93.4	78.0	64.7	69.9	42.2	28.0	**80.0**	73.8
YOLOv10 [34]	CSPDarkNet	24.4M	62.6	93.5	79.1	86.8	72.4	64.8	60.2	22.4	33.9	28.6	84.2
Defect detectors													
MSC-DNet [21]	ResNet50	34.1M	69.1	**97.7**	95.2	92.5	75.2	67.0	61.1	37.6	48.8	31.2	84.5
MSC-DNet [21]	ResNet101	-	71.6	95.5	96.1	94.9	76.5	66.5	65.8	34.1	53.4	48.5	84.0
EFD-YOLOv4 [37]	CSPDarkNet	-	54.7	96.3	**98.0**	85.3	75.0	53.3	43.2	18.2	50.0	27.3	0
DCC-CenterNet [23]	ResNet50	32.8M	61.9	84.1	85.5	96.2	77.3	50.9	54.8	30.2	13.9	49.9	76.6
Zhang's [38]	CSPDarkNet	-	71.9	97.8	88.9	96.2	79.1	67.3	53.2	33.2	43.7	75.8	83.6
Multi-branch detectors													
CFT [39]	CSPDarkNet	206.2M	74.4	96.0	78.5	90.1	76.5	68.4	67.9	49.8	**66.5**	72.8	77.1
ICAFusion [40]	CSPDarkNet	120.3M	72.5	91.8	81.6	90.9	72.2	65.1	72.0	54.8	54.1	65.5	77.6
SuperYOLO [41]	CSPDarkNet	9.9M	69.1	85.9	76.8	87.1	70.6	66.6	62.4	56.8	63.1	49.5	72.5
Ghost [42]	CSPDarkNet	7.7M	64.5	89.9	59.9	82.1	65.4	63.4	68.8	35.4	51.0	62.0	67.4
LAGNet (Ours)	MobileNetV3	31.9M	**76.6**	93.7	85.3	95.1	77.7	66.5	**75.0**	56.8	59.1	79.5	77.1
LAGNet-s (Ours)	MobileNetV3	8.9M	75.4	94.8	85.1	91.9	77.2	**70.5**	74.7	**60.1**	65.3	56.2	78.4

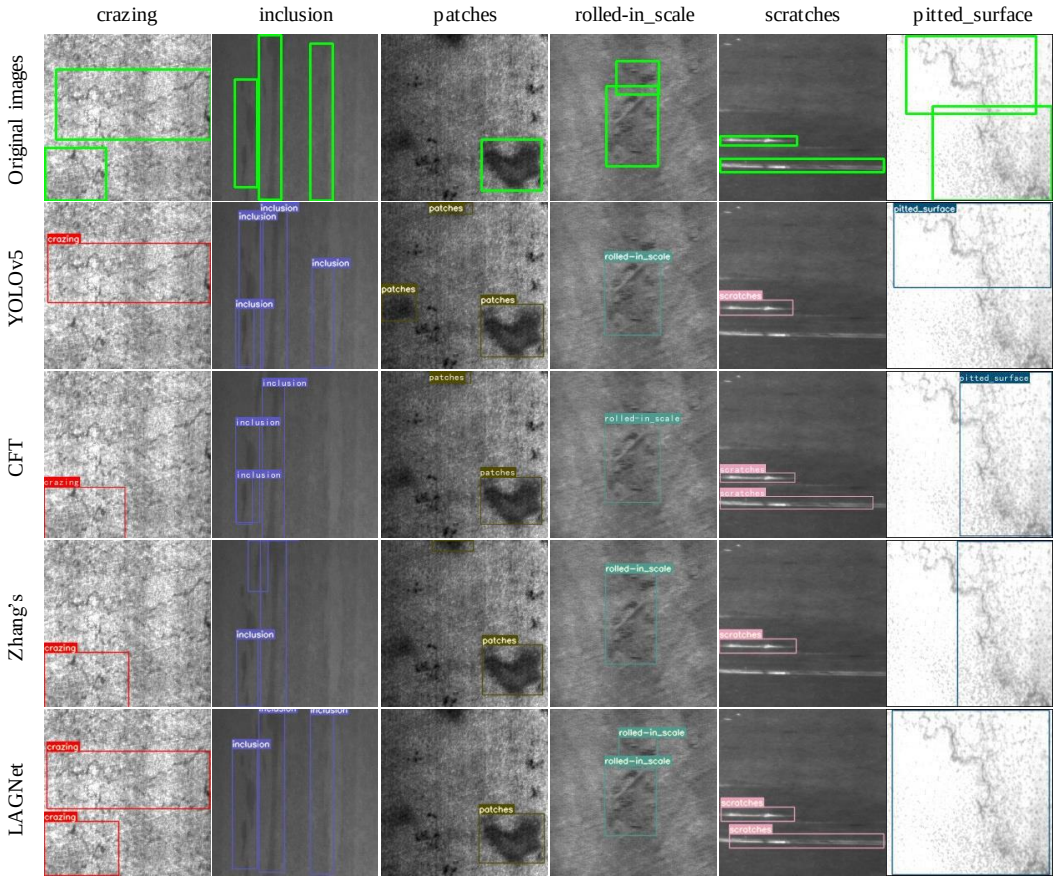


Figure 10. The visualization results of different models on NEU-DET. The green boxes in the first row indicate the ground-truth boxes.

other algorithms. Two-stage detectors such as Faster R-CNN (43.2%) and CascadeR-CNN (53.1%) do not perform well, mainly because they adopt a single-scale detection strategy and are not friendly to detect small defects in GB-DET. Especially, LAGNet also ranks the first AP in terms of 4 types of defects: *pit* (96.1%), *sand hole* (95.9%), *smudge* (81.8%), and *wrinkle* (97.1%). The results further validate the robustness and effectiveness of LAGNet.

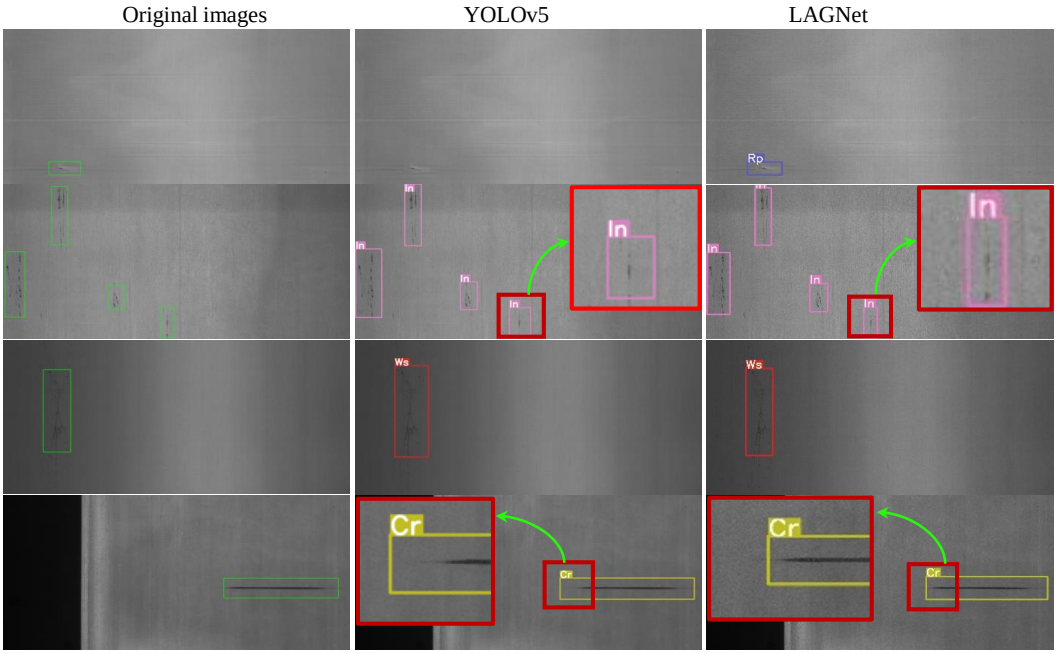


Figure 11. The visualization results of LAGNet on GC10-DET, where different colors represent different defect classes.

Table 3. Detection results of state-of-the-art methods on GB-DET.

Method	mAP(%)	R_angle	pit	sand_hole	smudge	wrinkle
Faster R-CNN [11]	43.2	90.1	72	3.9	50.2	0
Cascade R-CNN [44]	53.1	92.4	83.1	23.9	66.2	0
SSD512 [12]	64.0	91.6	78.1	34.6	55.3	60.4
RetinaNet [45]	72.1	83.2	66.2	62.9	63.2	84.9
YOLOv3 [15]	79.8	92.1	82.8	77.6	67.5	79.0
YOLOv5 [17]	90.6	94.6	91.5	91.8	78.1	97
YOLOX [33]	79.7	**96.2**	92.2	75.0	75.2	59.2
YOLOv7 [18]	88.8	94.9	90.2	90.6	73.8	94.7
SuperYOLO [41]	89.6	93.6	90.1	92.3	78.0	94.1
YOLOv10 [34]	89.3	93.7	94.8	92.2	78.0	87.6
LAGNet (Ours)	**93.2**	95.1	**96.1**	**95.9**	**81.8**	**97.1**

Table 4. Ablation studies of LAM and SCEM on NEU-DET.

Method	Backbone	Params	mAP(%)
YOLOv5	CSPDarkNet	46.1M	76.5
YOLOv5+LAM	CSPDarkNet	86.8M	82.3
YOLOv5+SCEM	CSPDarkNet	53.0M	78.6
YOLOv5+LAM+SCEM	CSPDarkNet	93.7M	84.6
YOLOv5+LAM	MobileNetV3	25.7M	83.6
YOLOv5+SCEM	MobileNetV3	28.6M	77.8
YOLOv5+LAM+SCEM	MobileNetV3	32.2M	**85.5**

4.5. Ablation studies

Impact of each module. We first validate the effects of the separate LAM branch and SCEM, as shown in Table 4. By incorporating the LAM branch into the baseline network using CSPDarkNet as the backbone, the mAP is dramatically improved from 76.5% to 82.3%. The significant performance gain results from the improvement of network localization capability, that is, the LAM could provide the RGB branch with more complementary location information. Meanwhile, integrating SCEM separately brings a 2.1-point performance gain as well. When both the LAM and SCEM are introduced, the mAP further goes up to 84.6%.

Considering the significant parameter increase when integrating both LAM and SCEM, we replace the backbone with the lightweight MobileNetV3. Surprisingly, not only the number of parameters drastically decreases to 32.2M from 93.7M, but the mAP also achieves an improvement of 0.9%. We argue that this may be attributed to the squeeze-and-excite bottleneck in MobileNetV3, which is beneficial for extracting features of weak defects. Additionally, the lightweight backbone's suitability for training defect datasets with fewer images may also contribute to this outcome.

Table 5. Ablation studies of kernel size and number in SCEM.

Dataset	mAP (n \ k)	1×3 3×1	1×5 5×1	1×7 7×1	1×9 9×1
NEU-DET	1	83.9	84.4	84.8	85.0
	2	84.0	**85.5**	84.5	83.8
	3	84.6	83.8	83.5	82.9
GC10-DET	1	71.3	74.1	73.4	**76.3**
	2	74.4	**76.6**	75.8	74.8
	2	74.3	70.7	70.2	68.0

Kernel size and number in SCEM. We conduct comparison experiments to validate the optimal kernel size (k) and number (n) of strip convolutions utilized in SCEM. As demonstrated in Table 5, the best results (85.5% mAP on NEU-DET and 76.6% mAP on GC10-DET) are achieved when applying two 1×5 and 5×1 strip convolutions in hor-

izontal and vertical directions, respectively. It is found that the performance can be first improved and then dropped as the kernel size and number increase. We believe that appropriately increasing the kernel size and number of strip convolutions helps enlarge the receptive field of neurons, but excessively large and many kernels may introduce irrelevant background interference information, thus hampering the detection accuracy. In conclusion, we set the kernel size $k = 5$ and number $n = 2$ in SCEM.

5. Conclusion

In this paper, we present a novel LAGNet for weak and strip surface defect detection. The LAM is introduced to focus on the discriminative areas and enhance the subtle visual differences between defective and defect-free areas. Meanwhile, the LGB inherently guides the RGB branch to pay more attention to the spatial features, thus improving the localization ability for weak defects in images. Furthermore, SCEM is delicately designed using depth strip convolutions, tailored for encoding the long-range spatial relations of strip defects efficiently. The experimental results demonstrate the effectiveness and efficiency of LAGNet. To further achieve efficient model deployment on industrial production lines, it is worth simplifying the network by getting rid of the auxiliary LAM branch in the inference stage in our future work.

Acknowledgement

This study was supported in part by the China Postdoctoral Science Foundation (Grant 2021TQ0301), and in part by the National Natural Science Foundation of China (Grant 62106232, Grant 62172371, Grant 62036010, and Grant U21B2037).

References

[1] J. Wang, Q. Li, J. Gan, H. Yu, and X. Yang. Surface defect detection via entity sparsity pursuit with intrinsic priors. *IEEE Transactions on Industrial Informatics*, 16(1):141–150, 2019. 1

[2] H. Wang, J. Zhang, Y. Tian, H. Chen, H. Sun, and K. Liu. A simple guidance template-based defect detection method for strip steel surfaces. *IEEE Transactions on Industrial Informatics*, 15(5):2798–2809, 2018. 1

[3] A. Zhou, H. Zheng, M. Li, and W. Shao. Defect inspection algorithm of metal surface based on machine vision. In *Proceedings of the IEEE international conference on Measuring Technology and Mechatronics Automation (ICMTMA)*, pages 45–49, 2020. 1

[4] S. Israni and S. Jain. Edge detection of license plate using sobel operator. In *Proceedings of the IEEE international conference on electrical, electronics, and optimization Techniques (ICEEOT)*, pages 3561–3563, 2016. 1

[5] Q. Zhou and H. Wang. CABF-YOLO: a precise and efficient deep learning method for defect detection on strip steel surface. *Pattern Analysis and Applications*, 27(2):36, 2024. 1

[6] K. Yang, Y. Liu, S. Zhang, and J. Cao. Surface defect detection of heat sink based on lightweight fully convolutional network. *IEEE Transactions on Instrumentation and Measurement*, 71:1–12, 2022. 1

[7] Y. Gong, M. Liu, X. Wang, C. Liu, and J. Hu. Few-shot defect detection using feature enhancement and image generation for manufacturing quality inspection. *Applied Intelligence*, 54:375–397, 2024. 1

[8] M. Everingham, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman. The pascal visual object classes (voc) challenge. *International Journal of Computer Vision*, 88(2):303–338, 2010. 1

[9] 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 (CVPR)*, pages 580–587, 2014. 2

[10] R. Girshick. Fast r-cnn. In *Proceedings of the IEEE international conference on computer vision (ICCV)*, pages 1440–1448, 2015. 2

[11] S. Ren, K. He, R. Girshick, and J. Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. *Advances in Neural Information Processing Systems*, 28, 2015. 2, 7, 9, 10

[12] W. Liu, D. Anguelov, D. Erhan, C. Szegedy, S. Reed, C.-Y. Fu, and A. C. Berg. Ssd: Single shot multibox detector. In *Proceedings of the IEEE European Conference on Computer Vision (ECCV)*, pages 21–37, 2016. 2, 7, 9, 10

[13] 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)*, pages 779–788, 2016. 2

[14] J. Redmon and A. Farhadi. Yolo9000: better, faster, stronger. In *Proceedings of the IEEE conference on computer vision and pattern recognition (CVPR)*, pages 7263–7271, 2017. 2

[15] J. Redmon and A. Farhadi. Yolov3: An incremental improvement. *arXiv preprint arXiv:1804.02767*, 2018. 2, 10

[16] A. Bochkovskiy, C.-Y. Wang, and H.-Y. M. Liao. Yolov4: Optimal speed and accuracy of object detection. *arXiv preprint arXiv:2004.10934*, 2020. 2

[17] G. Jocher. YOLOv5 by Ultralytics. https://github.com/ultralytics/yolov5, May 2020. 2, 7, 9, 10

[18] C.-Y. Wang, A. Bochkovskiy, and H.-Y. M. Liao. Yolov7: Trainable bag-of-freebies sets new state-of-the-art for real-time object detectors. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 7464–7475, 2023. 2, 7, 9, 10

[19] C. Zhao, X. Shu, X. Yan, X. Zuo, and F. Zhu. Rdd-yolo: A modified yolo for detection of steel surface defects. *Measurement*, 214:112776, 2023. 2

[20] S.-H. Gao, M.-M. Cheng, K. Zhao, X.-Y. Zhang, M.-H. Yang, and P. Torr. Res2net: A new multi-scale backbone architecture. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 43(2):652–662, 2019. 2

[21] R. Liu, M. Huang, Z. Gao, Z. Cao, and P. Cao. Msc-dnet: An efficient detector with multi-scale context for defect detection on strip steel surface. *Measurement*, 209:112467, 2023. 2, 9

[22] K. Duan, S. Bai, L. Xie, H. Qi, Q. Huang, and Q. Tian. Centernet: Keypoint triplets for object detection. In *Proceedings of the IEEE/CVF international conference on computer vision (ICCV)*, pages 6569–6578, 2019. 3

[23] R. Tian and M. Jia. Dcc-centernet: A rapid detection method for steel surface defects. *Measurement*, 187:110211, 2022. 3, 9

[24] X. Yu, W. Lyu, C. Wang, Q. Guo, D. Zhou, and W. Xu. Progressive refined redistribution pyramid network for defect detection in complex scenarios. *Knowledge-Based Systems*, 260:110176, 2023. 3

[25] X. Ding, Y. Guo, G. Ding, and J. Han. Acnet: Strengthening the kernel skeletons for powerful cnn via asymmetric convolution blocks. In *Proceedings of the IEEE/CVF international conference on computer vision (ICCV)*, pages 1911–1920, 2019. 3

[26] Q. Hou, L. Zhang, M.-M. Cheng, and J. Feng. Strip pooling: Rethinking spatial pooling for scene parsing. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (CVPR)*, pages 4003–4012, 2020. 3

[27] W. Yu, P. Zhou, S. Yan, and X. Wang. Inceptionnext: When inception meets convnext. *arXiv preprint arXiv:2303.16900*, 2023. 3

[28] M.-H. Guo, C.-Z. Lu, Q. Hou, Z. Liu, M.-M. Cheng, and S.-M. Hu. Segnext: Rethinking convolutional attention design for semantic segmentation. *Advances in Neural Information Processing Systems*, 35:1140–1156, 2022. 3

[29] A. Howard, M. Sandler, G. Chu, L.-C. Chen, B. Chen, M. Tan, W. Wang, Y. Zhu, R. Pang, V. Vasudevan, et al. Searching for mobilenetv3. In *Proceedings of the IEEE/CVF international conference on computer vision (ICCV)*, pages 1314–1324, 2019. 4

[30] Y. He, K. Song, Q. Meng, and Y. Yan. An end-to-end steel surface defect detection approach via fusing multiple hierarchical features. *IEEE Transactions on Instrumentation and Measurement*, 69(4):1493–1504, 2019. 6, 7

[31] X. Lv, F. Duan, J.-j. Jiang, X. Fu, and L. Gan. Deep metallic surface defect detection: The new benchmark and detection network. *Sensors*, 20(6):1562, 2020. 6

[32] X. Kou, S. Liu, K. Cheng, and Y. Qian. Development of a yolo-v3-based model for detecting defects on steel strip surface. *Measurement*, 182:109454, 2021. 6

[33] Z. Ge, S. Liu, F. Wang, Z. Li, and J. Sun. Yolox: Exceeding yolo series in 2021. *arXiv preprint arXiv:2107.08430*, 2021. 7, 9, 10

[34] A. Wang, H. Chen, L. Liu, K. Chen, Z. Lin, J. Han, and G. Ding. Yolov10: Real-time end-to-end object detection, 2024. 7, 9, 10

[35] R. Hao, B. Lu, Y. Cheng, X. Li, and B. Huang. A steel surface defect inspection approach towards smart industrial monitoring. *Journal of Intelligent Manufacturing*, 32:1833–1843, 2021. 7

[36] L. Wang, X. Liu, J. Ma, W. Su, and H. Li. Real-time steel surface defect detection with improved multi-scale yolo-v5. *Processes*, 11(5):1357, 2023. 7

[37] S. Li, F. Kong, R. Wang, T. Luo, and Z. Shi. Efd-yolov4: A steel surface defect detection network with encoder-decoder residual block and feature alignment module. *Measurement*, 220:113359, 2023. 7, 9

[38] L. Zhang, Z. Fu, H. Guo, Y. Sun, X. Li, and M. Xu. Multi-scale local and global feature fusion for the detection of steel surface defects. *Electronics*, 12(14):3090, 2023. 7, 9

[39] Q. Fang, D. Han, and Z. Wang. Cross-modality fusion transformer for multispectral object detection. *Available at SSRN 4227745*, 2022. 7, 9

[40] J. Shen, Y. Chen, Y. Liu, X. Zuo, H. Fan, and W. Yang. Icafusion: Iterative cross-attention guided feature fusion for multispectral object detection. *Pattern Recognition*, 145:109913, 2024. 7, 9

[41] J. Zhang, J. Lei, W. Xie, Z. Fang, Y. Li, and Q. Du. Superyolo: Super resolution assisted object detection in multimodal remote sensing imagery. *IEEE Transactions on Geoscience and Remote Sensing*, 61:1–15, 2023. 7, 9, 10

[42] J. Zhang, J. Lei, W. Xie, Y. Li, G. Yang, and X. Jia. Guided hybrid quantization for object detection in remote sensing imagery via one-to-one self-teaching. *IEEE Transactions on Geoscience and Remote Sensing*, 61, 2023. 7, 9

[43] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin. Attention is all you need. *Advances in Neural Information Processing Systems*, 30, 2017. 8

[44] Z. Cai and N. Vasconcelos. Cascade r-cnn: Delving into high quality object detection. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6154–6162, 2018. 10

[45] T.-Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollár. Focal loss for dense object detection. In *2017 IEEE International Conference on Computer Vision (ICCV)*, pages 2999–3007, 2017. 10