Pattern Recognition Letters 153 (2021) 118–125

Contents lists available at ScienceDirect

Pattern Recognition Letters

journal homepage: www.elsevier.com/locate/patrec

HMFCA-Net: Hierarchical multi-frequency based Channel attention net for mobile phone surface defect detection☆

Ying Zhu[a], Runwei Ding[a,*], Weibo Huang[a], Peng Wei[a], Ge Yang[a], Yong Wang[b]

[a] Key Laboratory of Machine Perception, Shenzhen Graduate School, Peking University, China
[b] School of AI, Chongqing University of Technology, China

ARTICLE INFO

Article history:
Received 23 June 2021
Revised 1 November 2021
Accepted 30 November 2021
Available online 2 December 2021

Edited by: Jiwen Lu

Keywords:
Defect detection
HMFCA-Net
Multi-frequency channel information
Local cross-channel interaction

ABSTRACT

The surface defect detection is an important process in the production of mobile phones. To detect various mobile phone surface defects and acquire detailed features of tiny defects, this paper proposes a Hierarchical Multi-Frequency based Channel Attention Net (HMFCA-Net). In particular, an attention mechanism that uses multi-frequency information and local cross-channel interaction is proposed to represent the weighted defect features. A deformable convolution based ResNeSt network is introduced to handle various defect shapes. Besides, to overcome the extreme aspect ratio problem caused by the tiny phone surface defects, a RoI Align is introduced to decrease localization error. Experiments on the public DAGM dataset and a self-collected dataset named MPSSD shows that the proposed method achieves promising performance on defect detection task.

1. Introduction

In the actual production process, the defect detection takes high labor costs in most mobile phone foundries since the existing manual or online sampling defect methods require a lot of labor. In addition, the subjective differences of individuals make it easy to cause false detections and missed detections. As a result, seeking an efficient, reliable, and accurate intelligent detection system to replace manual detection is of great significance for the quality control of products.

Due to the excellent robustness and high accuracy of deep neural networks, deep learning-based object detection techniques have been widely used in pedestrian detection [16], text detection [11], video surveillance [14,22], autonomous driving [4,5], defect detection [10,25,31], etc. In literature, there are some works that use deep learning-based methods to perform the surface defect detection of mobile phones [13,15,19]. Although significant progress has been achieved, most of the current methods directly apply deep learning-based object detection technologies to the defect detection domain, while they ignore that defect detection is somewhat different from object detection as the size of defects is usually extremely small compared to the whole image (i.e., the extreme aspect ratio problem) and the defect detection is easily affected by low contrast, background, etc. The first row of Fig. 1 shows some mobile phone surface defect examples, in which the defect areas are indicated by thick outlines. The defects are usually small compared to the whole image. (j)–(l) shows the extreme aspect ratio problem of mobile phone scratches. The aspect ratio between the width and height of scratches is extremely small, which is easy to cause localization error during regression process. Most of the existing methods extract the features on the whole image, which may contain redundant background information while contains little information about defects. To solve the problem, this paper pays more attention on the defect-related regions. In specific, a channel attention mechanism that uses multi-frequency information and local cross-channel interaction is firstly introduced to get detailed features of tiny defects. It's significant to accurately extract features related to defects in an image. This paper focuses on two concerns:

(1) **How to get more targeted defect features?** Most of the current deep learning-based methods extract features on the whole image. As Fig. 1 shows, (d) is the input image with a tiny defect, (e) shows the feature extracted by Faster R-CNN, which ignores the defect information and involves redundant background information. While (f) successfully extracts the defect region features, showing the priority of attention mechanism in extracting task-specific features. Attention needs to be paid to defects-related regions in mobile phone surface defect detection.
(2) **How to tackle the extreme aspect ratio problem and detect defects of different shapes?** As Fig. 1 shows, defects in the

☆ Editor: Emmanouil Benetos

* Corresponding author.
E-mail address: dingrunwei@pku.edu.cn (R. Ding).

https://doi.org/10.1016/j.patrec.2021.11.029
0167-8655/

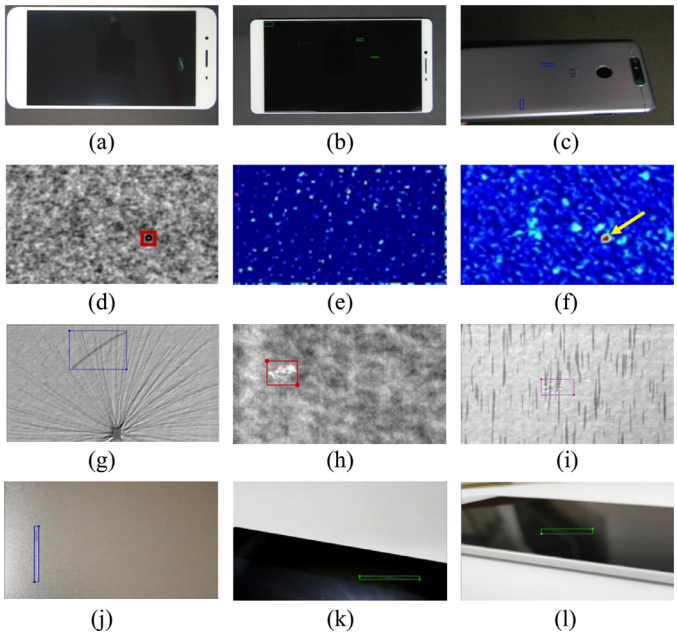

Fig. 1. The first row shows the mobile phone screen scratch defects and cover scratch defects. The second row illustrates the necessity of attention mechanism. (d) is the input image with a tiny defect, (e) shows the feature on the whole image extracted by Faster R-CNN, which ignores the tiny defect information. (f) shows the feature extracted by the proposed HMFCA-Net, which successfully acquires the defect feature. (g)–(i) are various defect shapes. (j)–(l) show the extreme aspect ratio problem of mobile phone scratches.

production line usually have various shapes and some of them even have extreme aspect ratio, which causes the defect localization error. It's urgent to find a solution to detect defects of different shapes and different aspect ratios.

Considering the foregoing concerns, this paper adopts attention mechanism and proposes a HMFCA-Net to extract and represent the defect features. The overview of the proposed framework is shown in Fig. 2. The deformable convolution-based network extracts feature of the input image, and features from Res2 to Res5 layer are further processed by HMFCA-Net. Then the HMFCA-Net enhances the defect feature representation of input features. It performs Discrete Cosine Transformation (DCT) on different frequency components and uses a local cross-channel interaction to keep the straight correspondence between channels and their weights. The FPN module performs down-sampling operation to combine the outputs of HMFCA-Net. Following the FPN module, the RoI Align module pools the combined feature into a fixed size for further classification and regression. The proposed network outputs the defect type and location at last.

The main contributions of this paper can be summarized as follows:

(1) This paper proposes a hierarchical multi-frequency based channel attention network (HMFCA-Net), which pays more attention on features related to defects through multi-frequency information and local cross channel interaction.
(2) This paper proposes an improved two-stage detector for mobile phone surface defect detection, where a deformable convolution-based ResNeSt network is proposed to detect defects with various shapes, and a RoI Align module is introduced to tackle the extreme aspect ratio problem.
(3) A self-collected MPSSD dataset containing 2644 images with two different scratches is produced and will be publicly available. Extensive experiments carried out on MPSSD and the public DAGM [27] dataset demonstrate the effectiveness of our method.

2. Related work

2.1. Surface defect detection methods

In addition to traditional manual detection methods, the surface defect detection of mobile phones can be divided into methods based on traditional machine vision and deep learning. For traditional machine vision learning-based methods, Jian et al. [12] used a vision-based defect detection system to automatically detect defects of mobile phone screens, in which a contour-based image registration algorithm was proposed to solve the misalignments in images caused by rotation and displacement. Weimer et al. presented a machine vision system using basic patch statistics combined with a two layer neural network to detect surface defects on arbitary textured and weakly labeled images [26]. The traditional machine vision-based methods usually rely on hand-craft features, which can be extracted by the statistical rules, mathematical principles, or user preferences. Although the hand-craft features are intuitive, they are still sensitive to image noise and application scenarios to some extent. Since the update cycle of mobile phones is very short, such as quarterly or half a year, traditional methods cannot adapt to the rapid adjustment of mobile phone production. To solve this problem, many researchers try to introduce deep learning methods into defect detection. For example, Lu et al. [19] proposed a generating sample method based on deep convolution generation antagonism network (DCGAN) for detecting the defects on mobile phone protection screen. Jie et al. [15] proposed an end-to-end screen defect detection network for mobile screen defect detection, in which the merging and splitting strategies were used to cope with multiple size and shape variations of defects. For detecting the defects on the back glass of mobile phones, a symmetric convolutional neural network composing of encoder and decoder structures was introduced by Jiang et al. [13]. For detecting the defects on the cover glass of mobile phones, Yuan et al. [30] proposed a modified segmentation method, where a data generation algorithm that combined with an augmentation process was presented to avoid the huge labelled data requirement. Different from the aforementioned methods, this paper focuses on getting more targeted defect features. HMFCA-Net is proposed to assign different weights to feature channels adaptively so that the channels relevant to defects are enhanced and redundant background information is weakened. Besides, a deformable convolution-based ResNeSt network and a RoI Align module are introduced to detect defects with various shapes and deal with extreme aspect ratio problem.

2.2. Attention mechanisms in CNN

Recently, there are several attempts to improve the performance of neural networks for vision-related tasks. Attention mechanism has been proven helpful for enhancing deep CNNs. Channel attention and spatial attention are two mainstream attention mechanisms. Channel attention pays different attention to different feature channels, while spatial attention pays different attention to different positions on the feature map. Channel attention was the core idea proposed by SE-Net [9]. Models can acquire features of different channels with different weights. The more important the channel is for the tasks, the bigger the relevant weight is. Woo et al. [28] proposed a convolutional block attention module (CBAM) to infer feature maps along channel and spatial dimensions, and the attention map was multiplied by the original input for feature refinement. In CBAM, global average pooling and max pooling were both used to get better information of inputs. Many

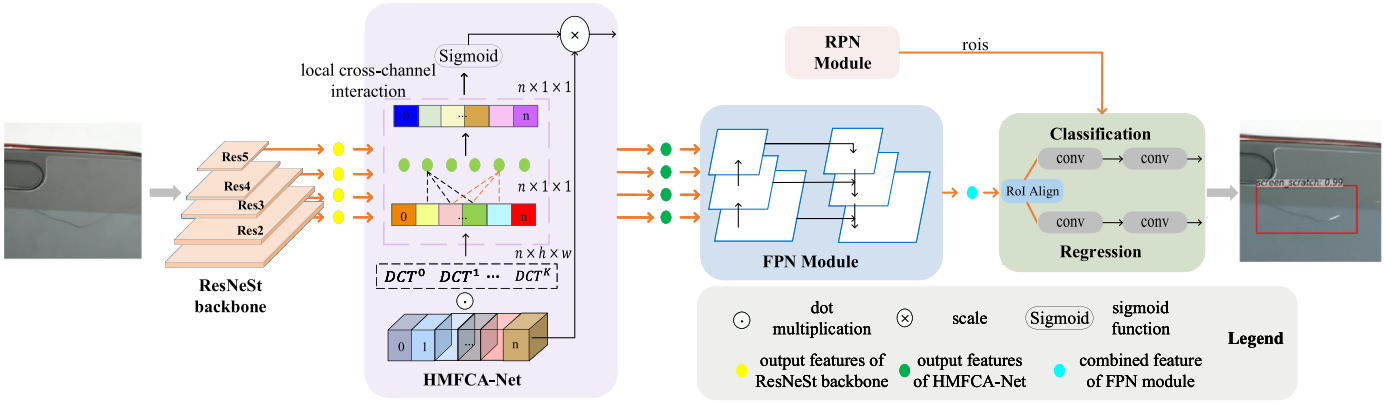

Fig. 2. Overview of our proposed method for mobile phone surface defect detection. The features from Res2 layer to Res5 layer of ResNeSt network are processed by HMFCA-Net. HMFCA-Net contains DCT transformation and local cross-channel interaction. FPN module combines the output features and the combined feature is used for further classification and regression.

improvements were made based on channel and spatial attention mechanisms [3,8]. ECA-Net [24] replaced the fully connected layers in channel attention with one-dimensional convolution and decreased model complexity. FcaNet [20] proposed a new preprocessing of channel attention to replace global average pooling. This work proved the insufficiency of global average pooling for rich feature representation. Fu et al. [6] proposed a dual attention network (DANet) which captured long-range contextual information in spatial and channel dimensions respectively and aggregated the outputs for pixel-level prediction. Zhang et al. [33] proposed deep residual channel attentions (RCAN) where residual channel attention was used to rescale channel-wise features.

In traditional channel attention mechanisms, global average pooling operation is generally used to get the input information on each channel, and two fully connected layers are used to get the weights of different channels. Despite the simplicity of global average pooling, it is inadequate to get rich information of inputs because various inputs can have the same mean value. Two fully connected layers reduce data dimensionality but destroy the direct correspondence between channels and their weights [24]. This paper takes advantage of FcaNet [20] and ECA-Net [24] and proposes a hierarchical multi-frequency based channel attention network (HMFCA-Net). Multi-frequency information of features and local cross-channel interaction are used to get weighted defect features, where the features of defect regions are enhanced. The outputs of HMFCA-Net are processed by feature pyramid networks to combine features in deep and shallow layers for better defect detection.

3. Method

Taking faster R-CNN [21] with FPN [17] as the baseline, this paper makes three improvements: HMFCA-Net, deformable convolution-based ResNeSt network, and RoI Align. The structure of the proposed method is shown in Fig. 2. This section will introduce it in detail.

3.1. Hierarchical multi-frequency based channel attention network

For simplicity, channel attention mechanism generally uses global average pooling to get initial weights of features for each channel, which calculates the mean value of all pixel points for each feature map. However, different features can share the same mean value after global average pooling operation, making them indistinguishable. Hence, it's unsuitable to apply global average pooling for tiny features in defect detection. Besides, it has been proven in Qin et al. [20] that global average pooling can be regarded as the lowest frequency component of two-dimension discrete cosine transform (2D DCT). It's redundant to extract information from channels at the same frequency component. Motivated by the above, this paper proposes HMFCA-Net combined with multi-frequency information and local cross-channel interaction to get weighted defect features, where features of defect regions are enhanced.

The flowchart of HMFCA-Net is shown in Fig. 2, in which the traditional global average pooling is taken place by the proposed multi-frequency information module. Denoting the features of Res2 to Res5 layers from ResNeSt network as χ_m, $m\epsilon\{1,2,3,4\}$ and the relevant output features of HMFCA-Net as o_m. The goal of the network is to learn a non-linear mapping $\phi : \chi_m \rightarrow o_m$. In multi-frequency information module, the input feature information on K frequency components at channel dimension is extracted through two-dimension discrete cosine transformation, which can be calculated as follows:

$$f(u,v) = \sum_{i=0}^{H}\sum_{j=0}^{W} x_{i,j} \cos\left(\frac{\pi u}{H}\left(\frac{1}{2}+i\right)\right)\cos\left(\frac{\pi v}{W}\left(\frac{1}{2}+j\right)\right) \quad (1)$$

$$s.t. \quad u\epsilon\{0,1,\ldots,H-1\}, \quad v\epsilon\{0,1,\ldots,W-1\},$$

where $f(u,v)$ is the frequency component of 2D DCT frequency spectrum, H and W are height and width of the input. In the following, the DCT weight term is denoted as $B^{i,j}_{u,v}$ for convenience:

$$B^{i,j}_{u,v} = \cos\left(\frac{\pi u}{H}\left(\frac{1}{2}+i\right)\right)\cos\left(\frac{\pi v}{W}\left(\frac{1}{2}+j\right)\right). \quad (2)$$

In HMFCA-Net, the input feature is divided into K (K equals to 16 in this work) parts along channel dimension. Denoting these parts as $\left[X^0, X^1, \ldots, X^K\right]$, each part is assigned by its corresponding 2D DCT frequency component. To get DCT weights of the input χ_i, the features of Res2 to Res5 layers are resized into 56×56, 28×28, 14×14, 7×7 at first. Then the DCT weights are constants. The output of 2D DCT is denoted as $F(X)$ and it is the concatenation of $F^m_{(u,v)}$, which can be calculated as follows:

$$F^m_{(u,v)} = \sum_{i=0}^{H}\sum_{j=0}^{W} x^m_{i,j} B^{i,j}_{u,v} \quad s.t. \quad m\epsilon\{0,1,\ldots,K-1\}. \quad (3)$$

HMFCA-Net applies $F(X)$ to replace global average pooling. After 2D DCT operation, local cross-channel interaction is considered to substitute traditional fully connected layers in channel attention. Two fully connected layers are used for dimension reduction in channel attention mechanism, which destroys the straight correspondence between channels and their weights and decreases the

performance of channel attention. Hence, HMFCA-Net takes local cross-channel interaction instead, and the structure of local cross-channel interaction is shown in Fig. 2. Denoting the output vector of 2D DCT as $y \epsilon R^C$ and the channel number of inputs as C, the equation of local cross-channel interaction can be computed as follows:

$$I(y) = W_g y, \tag{4}$$

$$W_g = \begin{bmatrix} w^1 & \dots & w^g & 0 & 0 & \dots & \dots & 0 \\ 0 & w^1 & \dots & w^g & 0 & 0 & \dots & 0 \\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ 0 & \dots & \dots & 0 & 0 & w^1 & \dots & w^g \end{bmatrix}, \tag{5}$$

where W_g is the weight matrix of local cross-channel interaction. All channels share the same learning parameters, so the local cross-channel interaction only involves g parameters. In this work, we empirically set $g = 3$. The output weight of each channel I^i is only related to g adjacent channels, as follows:

$$I^i = \sum_{j=0}^{g} w^j y_i^j \quad s.t. \quad i \epsilon \{0, 1, \dots, C\}, \tag{6}$$

where y_i^j indicates the jth adjacent channel of y_i. The sigmoid function is done after local cross-channel interaction. Hence, the whole process of HMFCA-Net can be calculated as follows:

$$\chi \to o : o_m = \chi_m \otimes sigmoid(I(F(\chi_m))), m \epsilon \{1, 2, 3, 4\}. \tag{7}$$

In HMFCA-Net, multi-frequency information and local cross-channel interaction are applied to get detailed features of tiny defects. Through training, more attention can be paid on defect-related regions.

3.2. Deformable convolution-based ResNeSt network

The proposed method takes ResNeSt network [32] as the feature map extractor. Deformable convolution is used to detect defects with various shapes. In this paper, the 3×3 convolution operations in the last three stages of ResNeSt network are substituted by 3×3 deformable convolution.

ResNeSt is a variant of ResNet and is a split attention network. Compared to ResNet, ResNeSt enables attention across feature-map groups, and the architecture requires for less computational cost and is easier to be applied to other algorithms. The stacking of deformable convolution-based ResNeSt blocks constitutes deformable convolution-based ResNeSt network. The structure of deformable convolution-based ResNeSt block is shown in Fig. 3. It shows that the input feature is divided into K groups for further operation, which are called as cardinal groups. There are R splits in a cardinal group, hence the total number of feature group is $G = KR$. After transformations $\{F_1, F_2, \dots, F_G\}$, the intermediate representation of each feature group U_i equals $F_i(x)$, $i \epsilon \{1, 2, \dots, G\}$. Then the kth cardinal group $\hat{U}_k$ is calculated as follows:

$$\hat{U}_k = \sum_{j=R(k-1)+1}^{Rk} U_j, \tag{8}$$

where the size of $\hat{U}_k$ equals $H \times W \times C/K$, with H, W, C respectively denote the height, width, and channel of the block output. Through the split attention in each cardinal group, we can get a weighted fusion of a cardinal group representation:

$$V_k = \begin{cases} \sum_{i=1}^{R} \frac{\exp(\vartheta_i(s_k))}{\sum_{j=0}^{R} \exp(\vartheta_j(s_k))} U_{R(k-1)+i} & , \text{if} \quad R > 1 \\ \sum_{i=1}^{R} \frac{U_{R(k-1)+i}}{1+\exp(-\vartheta_i(s_k))} & , \text{if} \quad R = 1 \end{cases} \tag{9}$$

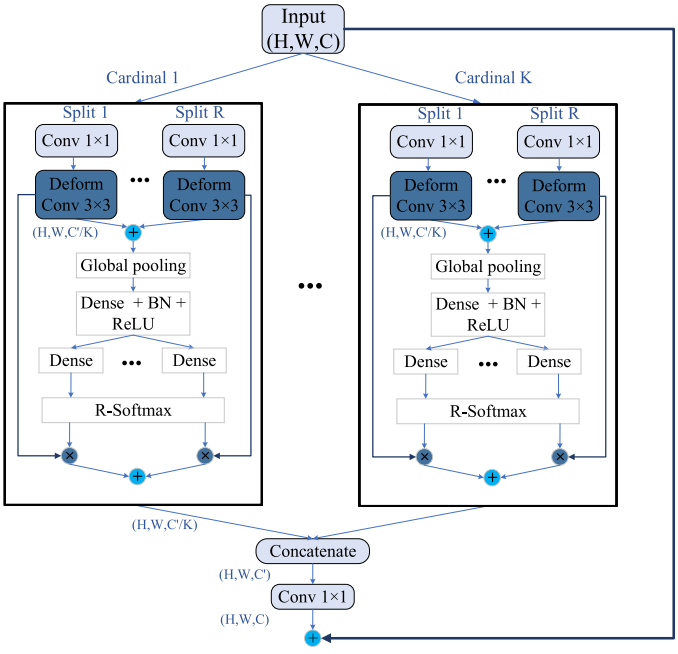

Fig. 3. The structure of deformable convolution based ResNeSt block.

Fig. 4. The comparison between traditional convolution (left) and deformable convolution(right). Compared to traditional convolution, it's easier for deformable convolution to extract the shape characteristic of defects.

where ϑ_i determines the weight of each split based on the global context representation s_k. s_k is calculated as follows:

$$s_k = \frac{\sum_{i=1}^{H} \sum_{j=1}^{W} \hat{U}_k(i, j)}{H \times W}. \tag{10}$$

As for deformable convolution, the 3×3 convolution operations in the last three stages of ResNeSt network are substituted by 3×3 deformable convolution. The comparison between traditional convolution and deformable convolution is shown in Fig. 4. The convolution position of deformable convolution is deformable according to offsets, while traditional convolution can only extract the features of the rectangular box. It is easier for deformable convolution to extract features of defects with different shapes. In summary, the proposed deformable convolution-based ResNeSt network combines the advantages of ResNeSt and deformable convolution, which is suitable for small defect detection with different shapes.

3.3. RoI align

For slender scratches with extreme aspect ratio in mobile phones, the localization error caused by the quantization process of RoI pooling must be eliminated. A tiny bias in quantization may lead to several pixels bias in the original image, which may make

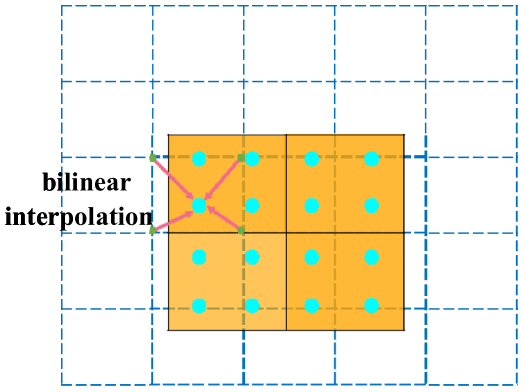

Fig. 5. The process of RoI Align. The white region is the whole input feature map and the yellow region is the corresponding bounding box region. RoI Align aims to resize the bounding box region into a fixed size $m \times m$. Divide the bounding box region into $m \times m$ parts at first. Then four fixed pixels (colored in cyan) are chosen to calculate values, and the value of each point is bilinear interpolated using the nearest four real pixels (colored in green).

Table 1
The distribution of defects in train, val, test sets.

	Train	Val	Test
screen_scratch	852	235	116
cover_scratch	998	293	150
total	1850	528	266

the model detect a wrong position for slender defects with extreme aspect ratio. To avoid the quantization process and improve the localization accuracy, a RoI Align module is applied.

Fig. 5 demonstrates the process of RoI Align. The white region is the whole feature map, and the yellow region is the corresponding bounding box region. The goal is to get an output of bounding box region with $m \times m$ size. The first step is to divide the bounding box region into $m \times m$ parts. We call each part a bin. It's likely that the vertices will not fall on the real pixels after being divided equally. Then four fixed pixels (colored in cyan) are chosen to calculate values. The value of each point is bilinear interpolated using the values of the nearest four real pixels (colored in green). Taking the biggest value in each bin as the output value, the output size is $m \times m$. m is set to 2 in this paper.

RoI Align eliminates the two quantization process in RoI pooling and can improve the localization performance for mobile phone surface scratches with extreme aspect ratio.

4. Experiments

4.1. Datasets

MPSSD dataset There are few datasets for mobile phone surface defect detection task. Thus we propose a mobile phone surface scratch dataset, named MPSSD. It contains two kinds of defects: screen scratches and cover scratches. The length of scratches ranges from 2 mm to 8.3 mm, and the width is much smaller than 1 mm. Hence the extreme aspect ratio is a common problem for scratch defects and challenges the localization performance of detectors. In total, this dataset includes 2644 images. Some examples are shown in Fig. 1(a)–(c). The average pixel size of each image is around 4000 × 3000. To simulate a real product line, we don't crop images into small sizes. The distribution of the dataset is shown in Table 1, where 1850 images are randomly selected to train object detection models, 528 images are selected to evaluate the models, and 266 images are selected for testing.

Table 2
Backbone comparisons.

Backbone	#P	GFLOPS
VGG16	138.4 M	15.5
ResNet-50	25.6 M	4.12
ResNet-101	44.5 M	7.84
ResNeS-t50	27.6 M	5.36

DAGM dataset DAGM dataset is artificially generated by DAGM (German Association for Pattern Recognition) and GNNS (German Chapter of the European Neural Network Society). There are 10 kinds of defects and 2100 images. Among them, 1046 images are used for training, and 1054 images are used for testing. The image size of DAGM dataset is 512 × 512.

4.2. Implementation details

The experiments are implemented on one NVIDIA RTX 2080Ti GPU and based on the detectron2 object detection framework [29]. Stochastic gradient descent (SGD) with momentum 0.9 is used to update parameters. This paper uses a cosine learning rate schedule at the first ten epoch for warm-up. When the warm-up is finished, the learning rate is initialized as 0.002. After 40 epochs the learning rate decreases from 0.002 and the weight decay value is 0.0001. The batch size is set to 6 in this paper and we train 50 epochs for all datasets. The commonly used mean average precision (AP50), mean average recall (AR50) and F1 score are applied to evaluate the methods. Each model is run ten times and the results with mean and standard deviation are given.

4.3. Comparisons and discussions

Results on MPSSD Table 2 shows the number of parameters and FLOPS of different backbones. The amount of parameters refers to the total weight parameter of all the parameterized layers of the model. FLOPS is defined as floating point operations, and can be used to imply the complexity of models. One FLOPS can be defined as an addition and a multiplication. Table 2 illustrates that the ResNet network and its variants are less complicated and require less memory than VGG network [23]. The amount of parameters of ResNest-50 is similar to ResNet-50 while the FLOPS is a little bigger than ResNet-50, which shows that the complexity of ResNeSt-50 network is between ResNet-50 and ResNet-101.

Table 3 shows the comparisons on MPSSD dataset. Ours achieves the best performance even without deformable convolution based ResNeSt module, which proves the effectiveness of HMFCA-Net and RoI Align module for defect detection task. We can get a better performance with deformable convolution based ResNeSt module, which illustrates the superiority of deformable convolution based ResNeSt compared with ResNet on extracting tiny defect features.

We compare HMFCA-Net with SOTA attention methods [9,20,28] on MPSSD dataset. For fair comparison, we use Faster R-CNN as detector and ResNeSt-50 along with FPN as backbone network. The result is shown in Table 4, and Fig. 6 shows the error bar of related methods. Our HMFCA-Net outperforms FcaNet by 0.44% and outperforms SE block by 1.87% on F1 score, which proves that HMFCA-Net is better at finding important channels related with defect features.

Results on DAGM Table 5 shows the results on DAGM dataset. Ours (without dc* based ResNeSt) performs better than the baseline Faster R-CNN method and ours achieves the best performance with the addition of dc* based ResNeSt network. Besides, we evaluate the inference speed of different methods, where the inference speed of ours is slightly lower than EDD-Net [7]. However, ours

Table 3
Performance on MPSSD dataset. dc* represents the deformable convolution.

Model	Backbone	AP50(%)	AR50(%)	F1 Score(%)
One stage:				
YoLo v4 [1]	ResNet-50	67.1 ± 0.56	–	–
RetinaNet [18]	ResNet-50	75.40 ± 0.31	85.98 ± 0.43	80.26 ± 0.36
Two stage:				
Cascade R-CNN [2]	ResNet-50	76.80 ± 0.24	86.17 ± 0.31	81.34 ± 0.27
Faster R-CNN [21]	ResNet-50	77.09 ± 0.20	86.36 ±0.09	81.46± 0.15
Ours (without dc* based ResNeSt)	**ResNet-50**	**77.72 ± 0.15**	**86.74 ± 0.06**	**81.98 ± 0.11**
Ours	**ResNeSt-50**	**79.79 ± 0.10**	**88.12 ± 0.14**	**83.75 ± 0.12**

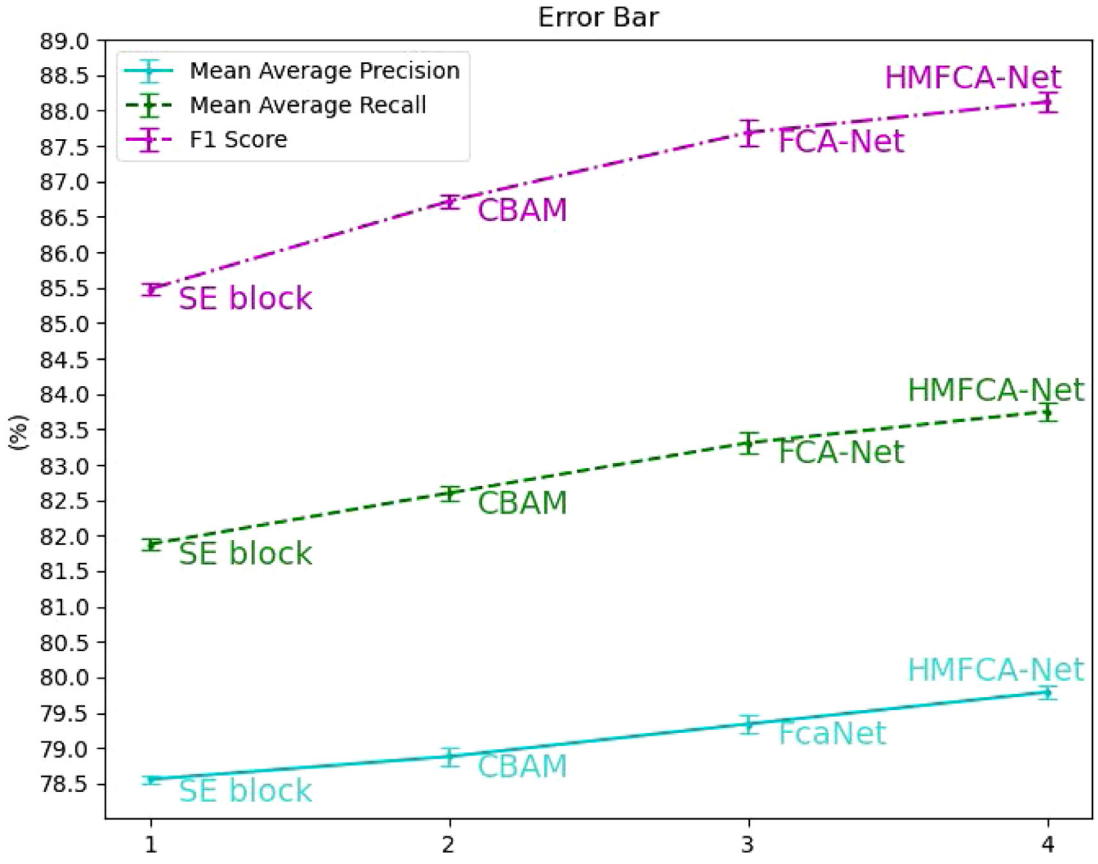

Fig. 6. The error bar of SE block[9], CBAM [28], FcaNet [20] and HMFCA-Net.

Table 4
Comparison of existing attention modules.

Methods	Detector	AP50 (%)	AR50(%)	F1 Score(%)
+ SE block [9]	Faster R-CNN	78.56 ± 0.06	85.48 ± 0.08	81.88 ± 0.07
+ CBAM [28]	Faster R-CNN	78.88 ± 0.13	86.72 ± 0.09	82.60 ± 0.11
+ FcaNet [20]	Faster R-CNN	79.34 ± 0.13	87.69 ± 0.18	83.31 ± 0.15
+ HMFCA-Net	Faster R-CNN	**79.79 ± 0.10**	**88.12 ± 0.14**	**83.75 ± 0.12**

achieves better performance at the same inference speed level. The proposed method can achieve the accuracy and speed request in industry product line.

Ablation study and qualitative visualization To demonstrate the effectiveness of HMFCA-Net, deformable convolution-based ResNeSt network and RoI Align, four ablation experiments are designed to compare with original Faster R-CNN using ResNet-50 and FPN as the backbone. The ablation study on MPSSD dataset is shown in Table 6. We can see that just changing the backbone to ResNeSt-50 improves performance, which shows the superiority of ResNeSt on extracting tiny defect features. With the help of deformable convolution, defects with various shapes are easier to be detected. Besides, it can be seen that the RoI Align can effectively improve the performance since it can decrease the localization error caused by extreme aspect ratio. HMFCA-Net helps the model to get more targeted defect features through weighing different channels adaptively. The proposed method (Faster R-CNN + deformable convolution-based ResNeSt network + HMFCA-Net + RoI Align) achieves the best performance.

Fig. 7 shows the extracted features on Res2 layer of different methods. Compared with features extracted by traditional Faster R-CNN and Cascade R-CNN method, HMFCA-Net pays more attention to defect-related regions and retains more defect information, which proves the effectiveness of HMFCA-Net for defect detection task.

Table 5
Performance on DAGM dataset.

Model	Backbone	AP50(%)	AR50(%)	F1 Score(%)	FPS
EDD-Net [7]	EfficientNet-B0	95.41 ± 0.37	98.53 ± 0.34	96.94 ± 0.35	33.5
EDD-Net [7]	EfficientNet-B1	97.14 ± 0.21	98.28 ± 0.25	97.70 ± 0.22	28.1
EDD-Net [7]	EfficientNet-B2	96.00 ± 0.23	97.25 ± 0.27	96.61 ± 0.26	24.3
Faster R-CNN [21]	ResNet-50	97.58 ± 0.17	98.23 ± 0.14	97.90 ± 0.15	28.8
Cascade R-CNN [2]	ResNet-50	97.68 ± 0.15	98.36 ± 0.21	98.02 ± 0.18	12.9
RetinaNet [18]	ResNet-50	97.85 ± 0.31	98.53 ± 0.19	98.17 ± 0.25	27.4
Ours(without dc* based ResNeSt)	ResNet-50	97.76 ± 0.15	98.50 ± 0.17	98.13 ± 0.16	22.9
Ours	**ResNeSt-50**	**98.17 ± 0.28**	**98.66 ± 0.23**	**98.41 ± 0.26**	19.2
Cascade R-CNN [2]	ResNet-101	97.74 ± 0.10	98.48 ± 0.09	98.11 ± 0.09	10.9
Faster R-CNN [21]	ResNet-101	97.90 ± 0.09	98.57 ± 0.05	98.24 ± 0.07	23.0
RetinaNet [18]	ResNet-101	98.00 ± 0.11	98.61 ± 0.03	98.30 ± 0.07	21.4
Ours (without dc* based ResNeSt)	**ResNet-101**	**98.79 ± 0.06**	**98.86 ± 0.07**	**98.82 ± 0.06**	18.3
Ours	**ResNeSt-101**	**98.85 ± 0.04**	**98.96 ± 0.06**	**98.90 ± 0.05**	14.1

Table 6
The ablation study.

Module	Faster R-CNN				
ResNeSt-50	✗	√	√	√	√
RoI Align	✗	✗	√	√	√
dc*-based ResNeSt-50	✗	✗	✗	√	√
HMFCA-Net	✗	✗	✗	✗	√
AP50	77.07	77.70	78.89	79.06	**79.79**

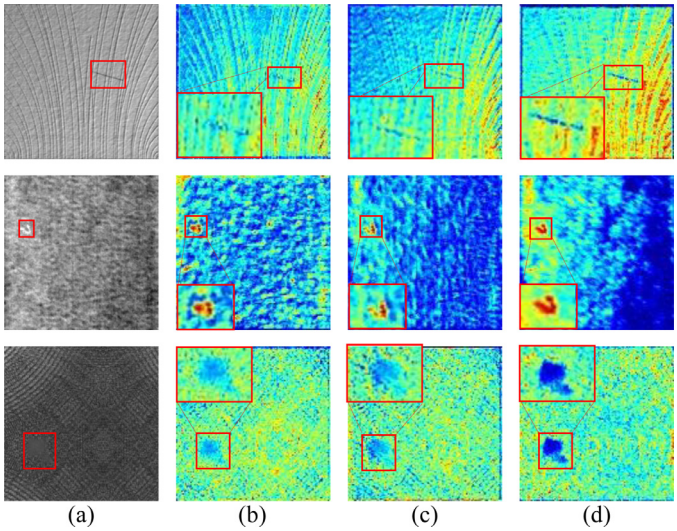

Fig. 7. Visualization comparisons of different models. The columns are respectively: (a) original images; (b) Features learned by Faster R-CNN; (c) Features learned by Cascade R-CNN; (d) Features learned by our HMFCA-Net.

5. Conclusion

This paper proposes HMFCA-Net to get detailed features of tiny defects, which pays more attention to defect-related regions through multi-frequency information and local cross-channel interaction. 2D DCT on several frequency components helps with getting more information about defects, and local cross-channel interaction keeps the straight correspondence between channels and their weights. Besides, deformable convolution-based ResNeSt network takes advantage of deformable convolution and ResNeSt network and helps the model to detect defects of various shapes. RoI Align avoids the quantization process in RoI pooling and decreases localization error of defects with extreme aspect ratio. This paper also proposes a mobile phone surface scratch dataset (MPSSD dataset). Extensive experiments on MPSSD dataset and DAGM dataset demonstrate the effectiveness of the proposed method. The proposed method achieves F1 score of 83.75% on MPSSD dataset and 98.41% on DAGM dataset, which performs best on both datasets.

Future work will include: (1) extending the proposed method to more types of mobile phone defects; (2) exploring to improve the inference speed.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgment

This work is supported by National Key R&D Program of China (2018YFB1308600, 2018YFB1308602).

References

[1] A. Bochkovskiy, C. Wang, H. M. Liao, Yolov4: optimal speed and accuracy of object detection, arXiv preprint arXiv:2004.10934
[2] Z. Cai, N. Vasconcelos, Cascade R-CNN: delving into high quality object detection, in: Proc. IEEE Conf. Comput. Vis., 2018, pp. 6154–6162, doi:10.1109/CVPR.2018.00644.
[3] L. Chen, H. Zhang, J. Xiao, L. Nie, J. Shao, W. Liu, T. Chua, SCA-CNN: spatial and channel-wise attention in convolutional networks for image captioning, in: Proc. IEEE Conf. Comput. Vis., 2017, pp. 5659–5667, doi:10.1109/CVPR.2017.667.
[4] X. Dai, HybridNet: a fast vehicle detection system for autonomous driving, Signal Process. Image Commun. 70 (2019) 79–88, doi:10.1016/j.image.2018.09.002.
[5] X. Dai, X. Yuan, L. Pei, X. Wei, Deeply supervised *z*-style residual network devotes to real-time environment perception for autonomous driving, IEEE Trans. Intell. Transp. Syst. 21 (6) (2019) 2396–2408, doi:10.1109/tits.2019.2918227.
[6] J. Fu, J. Liu, H. Tian, Y. Li, Y. Bao, Z. Fang, H. Lu, Dual attention network for scene segmentation, in: Proc. IEEE Conf. Comput. Vis., 2019, pp. 3146–3154, doi:10.1109/CVPR.2019.00326.
[7] T. Guo, L. Zhang, R. Ding, G. Yang, EDD-Net: an efficient defect detection network, in: Proc. Int. Conf. on Pattern Recog., 2021, pp. 8899–8905.
[8] J. Hu, L. Shen, S. Albanie, G. Sun, A. Vedaldi, Gather-excite: exploiting feature context in convolutional neural networks, in: 32nd Conference on Neural Information Processing Systems, 2018.
[9] J. Hu, L. Shen, G. Sun, Squeeze-and-excitation networks, in: Proc. IEEE Conf. Comput. Vis., 2018, pp. 7132–7141, doi:10.1109/TPAMI.2019.2913372.
[10] G. Hua, W. Huang, H. Liu, Accurate image registration method for PCB defects detection, J. Eng. 2018 (16) (2018) 1662–1667, doi:10.1049/joe.2018.8272.
[11] M. Jaderberg, K. Simonyan, A. Vedaldi, A. Zisserman, Reading text in the wild with convolutional neural networks, Int. J. Comput. Vis. 116 (1) (2016) 1–20, doi:10.1007/s11263-015-0823-z.
[12] C. Jian, J. Gao, Y. Ao, Automatic surface defect detection for mobile phone screen glass based on machine vision, Appl. Soft. Comput. 52 (2017) 348–358.
[13] J. Jiang, P. Cao, Z. Lu, W. Lou, Y. Yang, Surface defect detection for mobile phone back glass based on symmetric convolutional neural network deep learning, Appl. Sci. 10 (10) (2020) 3621, doi:10.3390/app10103621.
[14] A. Karpathy, G. Toderici, S. Shetty, T. Leung, R. Sukthankar, L. Fei-Fei, Large-scale video classification with convolutional neural networks, in: Proc. IEEE Conf. Comput. Vis., 2014, pp. 1725–1732, doi:10.1109/cvpr.2014.223.
[15] J. Lei, X. Gao, Z. Feng, H. Qiu, M. Song, Scale insensitive and focus driven mobile screen defect detection in industry, Neurocomputing 294 (2018) 72–81, doi:10.1016/j.neucom.2018.03.013.

[16] J. Li, X. Liang, S. Shen, T. Xu, J. Feng, S. Yan, Scale-aware fast R-CNN for pedestrian detection, IEEE Trans. Multimed. 20 (4) (2017) 985–996, doi:10.1109/tmm.2017.2759508.
[17] T. Lin, P. Dollár, R. Girshick, K. He, B. Hariharan, S. Belongie, Feature pyramid networks for object detection, in: Proc. IEEE Conf. Comput. Vis., 2017, pp. 2117–2125, doi:10.1109/CVPR.2017.106.
[18] T. Lin, P. Goyal, R. Girshick, K. He, P. Dollár, Focal loss for dense object detection, in: Proc. IEEE Int. Conf. Comput. Vis., 2017, pp. 2980–2988, doi:10.1109/TPAMI.2018.2858826.
[19] Y. Lu, L. Ma, H. Jiang, A light CNN model for defect detection of LCD, in: Proc. Int. Conf. Front. Comput., 2019, pp. 10–19. 978-981-15-3250-4_2
[20] Z. Qin, P. Zhang, F. Wu, X. Li, FcaNet: frequency channel attention networks, in: Proc. IEEE Conf. Comput. Vis., 2021, pp. 783–792.
[21] S. Ren, K. He, R. Girshick, J. Sun, Faster R-CNN: towards real-time object detection with region proposal networks, IEEE Trans. Pattern Anal. Mach. Intell. 39 (6) (2017) 1137–1149, doi:10.1109/TPAMI.2016.2577031.
[22] W. Shin, S. Bu, S. Cho, 3D-convolutional neural network with generative adversarial network and autoencoder for robust anomaly detection in video surveillance, Int. J. Neutal Syst. 30 (06) (2020) 2050034, doi:10.1142/s0129065720500343.
[23] K. Simonyan, A. Zisserman, Very deep convolutional networks for large-scale image recognition, Comput. Sci. (2014). arXiv preprint arXiv:1409.1556.
[24] Q. Wang, B. Wu, P. Zhu, P. Li, W. Zuo, Q. Hu, ECA-Net: efficient channel attention for deep convolutional neural networks, in: Proc. IEEE Conf. Comput. Vis., 2020, pp. 11534–11542, doi:10.1109/cvpr42600.2020.01155.
[25] P. Wei, C. Liu, M. Liu, Y. Gao, H. Liu, CNN-based reference comparison method for classifying bare PCB defects, J. Eng. 2018 (16) (2018) 1528–1533, doi:10.1049/joe.2018.8271.
[26] D. Weimer, H. Thamer, B. Scholz-Reiter, Learning defect classifiers for textured surfaces using neural networks and statistical feature representations, Procedia CIRP 7 (2013) 347–352.
[27] M. Wieler, T. Hahn, Weakly supervised learning for industrial optical inspection, 29th Annual Symposium of the German Association for Pattern Recognition, 2007.
[28] S. Woo, J. Park, J. Lee, I.S. Kweon, CBAM: convolutional block attention module, in: Proc. IEEE Eur. Conf. Comput. Vis., 2018, pp. 3–19, doi:10.1007/978-3-030-01234-2_1.
[29] Y. Wu, A. Kirillov, F. Massa, W. Lo, R. Girshick, Detectron2, 2019, (https://github.com/facebookresearch/detectron2).
[30] Z. Yuan, Z. Zhang, H. Su, L. Zhang, F. Shen, F. Zhang, Vision-based defect detection for mobile phone cover glass using deep neural networks, Int. J. Precis. Eng. Manuf. 19 (6) (2018) 801–810, doi:10.1007/s12541-018-0096-x.
[31] C. Zhang, W. Shi, X. Li, H. Zhang, H. Liu, Improved bare PCB defect detection approach based on deep feature learning, J. Eng. 2018 (16) (2018) 1415–1420, doi:10.1049/joe.2018.8275.
[32] H. Zhang, C. Wu, Z. Zhang, Y. Zhu, Z. Zhang, H. Lin, Y. Sun, T. He, J. Mueller, R. Manmatha, et al., ResNeSt: split-attention networks, arXiv preprint arXiv:2004.08955
[33] Y. Zhang, K. Li, K. Li, L. Wang, B. Zhong, Y. Fu, Image super-resolution using very deep residual channel attention networks, in: Proc. IEEE Eur. Conf. Comput. Vis., 2018, pp. 286–301, doi:10.1007/978-3-030-01234-2_18.