Neural Computing and Applications (2020) 32:11229–11244
https://doi.org/10.1007/s00521-020-04819-5

REVIEW ARTICLE

A robust weakly supervised learning of deep Conv-Nets for surface defect inspection

Haiyong Chen[1,2] · Qidi Hu[1] · Baoshuo Zhai[1] · He Chen[1] · Kun Liu[1]

Received: 4 September 2019 / Accepted: 24 February 2020 / Published online: 6 March 2020

Abstract

Automatic defect detection is a challenging task owing to the complex textured background with non-uniform intensity distribution, weak differences between defects and background, diversity of defect types, and high cost of annotated samples. In order to solve these challenges, this paper proposes a novel end-to-end defect classification and segmentation framework based on weakly supervised learning of a convolutional neural network (CNN) with attention architecture. Firstly, a novel end-to-end CNN architecture integrating the robust classifier and spatial attention module is proposed to enhance defect feature representation ability, which significantly improves the classification accuracy. Secondly, a new spatial attention class activation map (SA-CAM) is proposed to improve segmentation adaptability by generating more accurate heatmap. Moreover, for different surface texture, SA-CAM can significantly suppress the background's inference and highlight defect area. Finally, the proposed weakly supervised learning framework is trained using only global image labels and devoted to two main visual recognition tasks: defect samples classification and area segmentation. At the same time, it is robust to complex backgrounds. Results of the experiments verify the generalization of the proposed method on three distinct datasets with different kinds of textures and backgrounds. In the classification tasks, the proposed method improves accuracy by 0.66–25.50%. In the segmentation tasks, the proposed method improves accuracy by 5.49–7.07%.

Keywords Machine vision · Spatial attention · Deep learning · Defect detection · Convolutional neural network

1 Introduction

Surface defect inspection is important to production quality control in the intelligent manufacturing industry. Most surface defect inspection tasks in the manufacturing industry are still performed manually. Unfortunately, the disadvantages of manually defects inspection are obvious: subjective unstable and time-consuming. To overcome the disadvantages, automated surface inspection (ASI) technology is utilized to help or replace humans work. Among kinds of ASI methods, machine vision-based defect inspection methods have been wildly employed for surface quality controlling in manufacturing industry to help real-time identify and reject defective products, which can improve the production's quality and lifetime. Figure 1 shows three different typical defects on various kinds of surface texture. Moreover, different kinds of surface defects contain various features with random shapes and sizes in different directions and locations, which bring a huge challenge for visual defects inspection.

In fact, surface defects are local anomalies in various backgrounds. The existing surface defect algorithms mainly focus on the following four types of surfaces [1]: (1) non-textured surface; (2) repeated pattern surface; (3) homogeneously textured surface; and (4) non-homogeneously textured surface. Traditionally, automatic surface inspection methods are designed based on manually features. For non-texture surface, Luo et al. [2] used GC-LBP

✉ Haiyong Chen
haiyong.chen@hebut.edu.cn

Qidi Hu
marshellhu@qq.com

Kun Liu
liukun@hebut.edu.cn

[1] School of Artificial Intelligence and Data Science, Hebei University of Technology, Tianjin 300000, China

[2] State Key Laboratory for Reliability and Intelligence of Electrical Equipment, Hebei University of Technology, Tianjin 300000, China

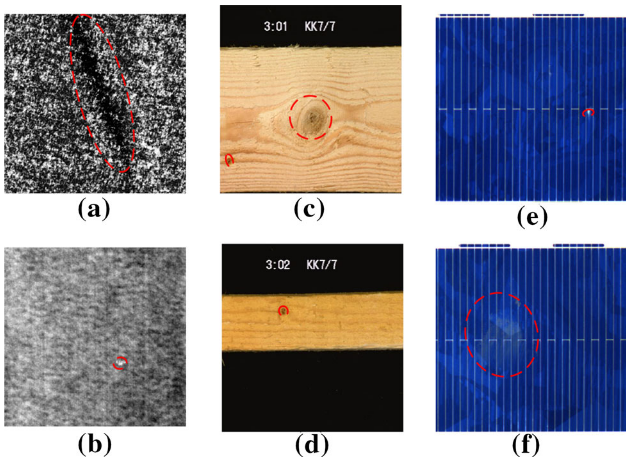


Fig. 1 Various types of texture defects. **a**, **b** Defects on a fabric surface. **c**, **d** Defects on a wood surface. **e**, **f** Defects on Polycrystalline silicon solar cell

to inspect steel surface defection. But this method might be not effective for complex backgrounds with random textures. In order to solve this problem, Su et al. [3] proposed a novel BCPICS-LBP descriptor. This new LBP-based method improved the crack defects recognition effect of different shapes and sizes compared with the traditional LBP features. The accuracy achieved 88.66%. However, the LBP feature is difficult to describe the surface defect characteristics of the repeat pattern surface because it does not consider the structural information of the repetitive texture. For repeated pattern surface images like fabric [3, 4], statistical representation like redundant contourlet transform (RCT) is used. The method's accuracy achieved 94.6%. As to homogeneously textured surface images such as wood surface, Wang et al. [5] used Gabor filters to learn the features of the wood surface and achieved a classification accuracy of 91.2%.

However, the manual feature's shortcomings for defect inspection were obvious: (1) sensitive to changes of the texture background and (2) highly dependent on human's knowledge. Moreover, it is difficult to construct feature represent for multiple surface defect and texture generally. These restrictions limited robustness and generalization of manual feature-extracting methods. So, automatically extracting feature methods are given more and more attention.

In recent years, deep learning (DL)-based methods are significantly researched. They are able to extract feature automatically and have been achieving good performance on image-related tasks [6–9] like classification and segmentation. For classification task, Jung [10] proposed a defect classification method for wood with randomly textured surfaces by employing different structures of CNNs. The method achieved accuracy for 92.30%. Chen et al. [11] proposed a multi-spectral CNN structure for classifying surface defects of solar cells. By separating the different channels of original image and separately convolution, the paper obtained 88.24% classification accuracy results, 2% higher than the traditional CNN. Zhou et al. [12] designed a CNN to learn multiple feature representations. The network classification accuracy achieved 97.3%. However, most of the above-mentioned CNN-based methods are designed to solve the problems for a specific texture surface. They are sensitive to minor changes in complex backgrounds. Moreover, the robustness of fully connected layer of original network may be weak for datasets with complex backgrounds. These CNN-based automatic feature-extracting methods achieve higher classification accuracy than traditional manual features.

To deal with the inner class separation and interclass compactness problem of softmax classifiers from original CNN, some papers redeveloped the structure with more robust classifiers. Tang et al. [13] replaced the fully connect (FC) layer with support vector machine (SVM). Based on three different common datasets, compared to traditional CNN structure, CNNs with SVM classifiers achieved 3% performance improvement, which verified the generality of this method. Further, Merentitis et al. [14] designed a combination of random forest and CNN on high-resolution remote sensing images. In the forest remote sensing dataset, this method's accuracy was 6% higher than the traditional CNN, whose increment was more significant than replacement of the full connectivity layer with SVM.

For ASI, after classifying the defect samples, it is necessary to segment the defect area to further find out the cause of the defect and troubleshoot the malfunction. Zhang et al. [15] designed a fabric defect detection framework based on YOLOv2. It could predict both the location and classification information of defect regions. The method achieved 69.45% intersection over union (IOU). Singh et al. [16] detected road damage and crack using mask-RCNN. Due to the complexity and randomness of the street and cement background, they only achieved 50% IOU for the road damage detection task. These supervised methods are trained with annotations.

Although the defect segmentation methods are effective, they require precise pixel-level annotation during training. However, because of the low occurrence of the defective sample and random changes from light intensity or complex backgrounds, it is extremely expensive to collect the accurately pixel annotated defect images. Therefore, the need for large amount of annotated data are still mainly weak points of CNN-based methods as Alan Yuille suggested [17].

To solve leakage problem of annotation defective samples in CNN-based methods, weakly supervised defect inspection has been extensively studied. Class activation map (CAM) [18] is one of commonly used inspection methods in CNN-based weakly supervised learning defect detection framework. Ren et al. [19] proposed a generic approach for automatic surface inspection in several image datasets with different kinds of texture background. Lin et al. [20] proposed a LED defect detection framework based on CAM for visual prediction of blocks. Li et al. [21] generated CAM for manipulation images to predict location of weak structure parts where collapse begins. In the above papers, the pixel or bounding box level inspection of defect is achieved by global image label, which avoids costly manual annotation.

Though weakly supervised methods reduce the requirement of label, the biggest problems of these methods are leakage of robustness. They are sensitive to background and texture's inference. In order to suppress interference from complex backgrounds, spatial attention mechanism is imported. According to [22], spatial attention mechanism improved the representation of interest area. Paper 6 [23] imported a saliency attention mechanism into CNN to detect the object, which improved the performance significantly compared with the standard FCN structure. Inspired by the above paper, we integrated CAM with attention mechanism and proposed spatial attention class activation map (SA-CAM). The proposed SA-CAM is able to focus on important features and suppress backgrounds' texture.

Comparing all the above-mentioned papers, a generic defect inspection framework based on CNN with attention architecture and random forest classifier is proposed in this paper. The contribution can be expressed as follows:

1. A novel deep CNN model is proposed for the defect classification problem in distinct surface textures by fusing CNN with random forest classier and spatial attention module. The random forest (RF) classifier is robust to changes in complex backgrounds. The spatial attention module can guide the CNN to gain more discriminating features. Thus, the novel CNN model significantly improves the classification effect and robustness of the proposed CNN model.
2. A robust spatial attention class activation map (SA-CAM) network structure is designed by integrating the above attention mechanism and CAM. The SA-CAM suppressed complex background with different textures and simultaneously highlight defective area, which is helpful to generate more accurate saliency map.
3. By depending the saliency map from SA-CAM, the proposed weakly supervised learning segmentation method uses global image label to achieve pixel-level defect segmentation, which simplifies the task of heavy pixel-level annotation for complex surface defects and unfolds good versatility for different texture surfaces.

The rest of the paper is organized as follows. In Sect. 2, the proposed framework containing random forest and spatial attention mechanisms is described. Section 3 presents the experimental results including classification, segmentation, and examples heatmap of some segmentation images. In Sect. 4, the results are further discussed and conclusion is given.

2 Proposed method

The proposed robust weakly supervised learning of deep Conv-Nets for surface defect inspections (RWSLDC) structure is shown in Fig. 2. The proposed framework includes feature extraction network, classification module, SA-CAM module, and segmentation module.

2.1 Feature extraction network

CNN is a type of feed-forward neural network. Generally, CNNs includes three major parts: (1) convolutional layer; (2) pooling layer; and (3) fully connected layer. The convolutional layer applies a group of convolutional filters on the local regions of the input, thus obtaining the feature maps of the input image. Suppose the number of filters is k, W_i denotes the weight of filter i, b_i is the bias of filter i, x_s stands for a small patch, σ is the activation function, and the size of input image is $a \times b$. The convolution of x_s given filter i is shown in

$$f(i, s) = \sigma(W_i x_s + b_i) \tag{1}$$

The pooling layer downsamples inputting spatial dimensions. For example, max pooling of image patch x_s is simply:

$$\text{pooling} = \max(x_s) \tag{2}$$

The pooling layer is normally applied after the convolutional layer to reduce the feature dimension and to avoid overfitting problem. For instance, the pooling with input size $a \times b$ and patch size $c \times d$ produces output with size $[(a-1)/c] \times [(b-1)/d]$, which is named of vision field. Generally, the pooling methods include average pooling, max pooling, and the Gaussian pooling. The fully connected layers normally constitute the last few layers of a CNN, whose works are computing the class scores and give out the classification results. A deep CNN normally consists of alternating convolutional and pooling layers, followed by fully connected layers. The CNN structure is widely used in computer vision-related tasks, such as object detection, scene classification, and video analysis. Among the various CNNs, the CNN model in Ref [11] is one of the effective models for surface defect inspection which contains five convolutional, three pooling, and three

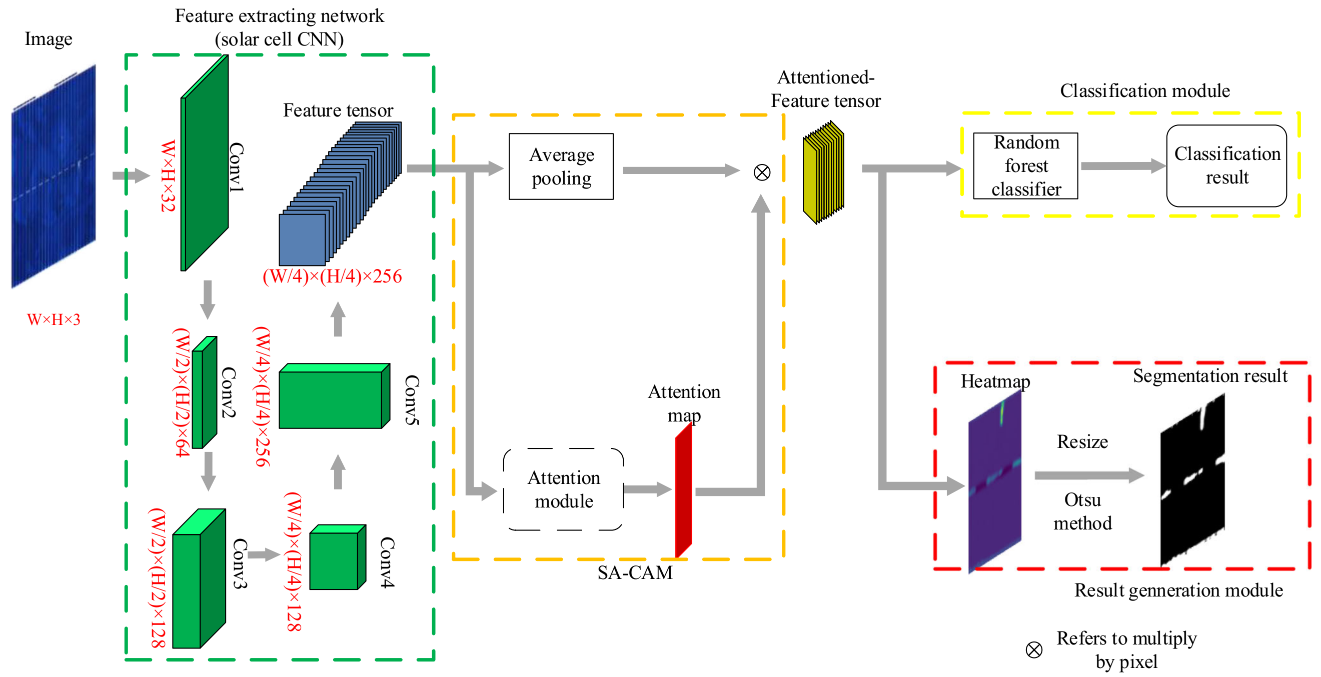


Fig. 2 RWLSDC framework structure

FC layers. The model structure of solar cell CNN references [11] and Table 1. It selects max pooling as down-sampling method. In this paper, the convolution layers will be used to extract features.

2.2 Classification module

In order to further improve the classification accuracy and adaptability of the proposed CNN, the classifier of the traditional convolutional neural network is redeveloped. Traditionally, the FC layers of CNN perform as the classifier. In this paper, a more robust random forest classifier is introduced into the CNN to replace the original fully connected layer.

Table 1 Structure of solar cells CNN

Name	Kernel	Solar cells CNN	
		Structures	Output
Layer1	16 × 7 × 7	Conv1	256 × 256 × 16
	2 × 2	Pool1	128 × 128 × 16
Layer2	32 × 5 × 5	Conv2	128 × 128 × 32
	32 × 5 × 5	Conv3	128 × 128 × 32
	2 × 2	Pool2	64 × 64 × 32
Layer3	64 × 3 × 3	Conv4	64 × 64 × 64
	64 × 3 × 3	Conv5	64 × 64 × 64
	2 × 2	Pool3	32 × 32 × 64
FC1	512	FC1	
FC2	512	FC2	
Softmax	2	Softmax	

The random forest algorithm [24] is a combined algorithm based on classification and regression decision trees proposed by Breiman et al. It is a classifier that uses multiple decision trees to train and predict samples classes. The construction of random forest model is generated by the following three steps:

1. Obtain the training dataset with N samples, randomly extract K samples by using the returning sampling, and then obtain K training subsets $\{D1, D2, \ldots, DK\}$.
2. The obtained training subset is used $(1 \leq i \leq K)$ to construct the sub-decision tree. There are M sample features and selecting F samples from M to form a random feature subspace as the split attribute set of the current node of the decision tree.
3. Every tree will grow equally without pruning. Finally, each tree will give out a result, and then vote of the decision trees are counted. The most votes are the output of the random forest.

Random forest classifier has following advantages: strong robustness, good generalization ability, and fast calculation speed. Based on the above advantages, it is selected as a classifier in the multiple distinct classification tasks, replacing the fully connected layer in the traditional convolutional neural network.

2.3 Spatial attention class activation map (SA-CAM)

Due to the randomness of surface defect locations, this paper merges the spatial attention into convolutional neural networks. Different from considering each image area equally, the spatial attention mechanism pays more attention to semantically related areas.

In the sample image of the defect detection dataset, the defect only accounts for a small portion of the entire image. In order to detect the defect area more accurately without reducing the picture resolution, it is necessary to separate the whole defect sample into smaller patches, which is called sliding window processing. In this case, pixel level is not directly available, and only patch-level results are accessed in the dataset. To extract the defect area, global average pooling (GAP) [25] can be used to extract the Class Activation Map (CAM) [18]. The highlighted part of the class activation map represents high possibility of defection.

CAM's calculating process is as follow: For a given image, let $f_k(x,y)$ represent the activation of sample k in the last convolutional layer at spatial location (x,y). Then, for the k, the result of performing global average pooling F_k is $\sum_{x,y} f_k(x,y)$. Thus, for a given class c, the input to the softmax, Sc, is $\sum_k w_c^k F_k$ where w_c^k is the weight of class c for unit k. Importantly,w_c^k refers to the importance of F_k for class c. Finally, the output of the softmax for the specific class is given by:

$$Pc = \frac{\exp(Sc)}{\sum_c \exp(Sc)}. \tag{3}$$

By applying $F_k = \sum_{x,y} f_k(x,y)$ the class score, we obtain Sc = $Sc = \sum_k \sum_{x,y} w_c^k f_k(x,y)$. Mc is defined as the class activation map for class c, where each spatial element is given by

$$Mc(x,y) = \sum_k w_c^k f_k(x,y) \tag{4}$$

In this paper, instead of using GAP, a trainable spatial attention method named spatial attention class activation map (SA-CAM) is proposed. It consists of two successive fully connected 1×1 convoluting layer which takes an instance feature. For the $Mc(x,y)$ in (3), the proposed spatial attention structure calculation formula is given by

$$Mc(x,y) = \sum_k w_{(x,y)} w_c^k f_k(x,y). \tag{5}$$

Here $w_{x,y}$ represents the attention weight. It takes a high-level feature vector $Mc(x,y)$ as the input and outputs an attention weight $w_{x,y}$ which is given by:

$$w_{(w,y)} = \mathrm{softmax}(w_2\ \mathrm{relu}(w_1 f_k(x,y) + b)) \tag{6}$$

where $w_2 \in R^L$ and $w_1 \in R^{L \times c}$ are trainable weight parameter matrices of two layers of the attention network, $b \in R^L$ is the bias parameter matrix, the structure is showed in Fig. 3.

2.4 Segmentation module

After passing through the GAP layer and generating the CAM, the generated saliency map is a heat value map. In order to obtain an accurate defect area, the resulting heat value map needs to be processed into a binary image. Commonly used image binarization methods are grayscale averages, fixed thresholds, the Otsu [26] method, and so on. Among these methods, the Otsu method (maximum interclass variance) can minimize the probability of pixel misclassification during grayscale binarization, so it is used as the thresholding method. It divides the image into background and target according to the grayscale characteristics of the image. The larger the variance between the background and the target, the greater the difference between the two parts that make up the image. When the partial target is divided into the background or the partial background is divided into the target, the difference between the two parts will be reduced. Therefore, the segmentation that maximizes the variance between classes means that the probability of misclassification is minimal.

Since all the original images' sizes are too large to detect the small defection, sliding windows detection is used to crop the image into small patches. Each patch is then individually fed into the trained convolutional neural network for identification and the corresponding heatmap results are output. If an image patch is identified as a defect-free picture, the heatmap value of the entire patch is forced to zero. Finally, all small patches are stitched together to form a complete heatmap.

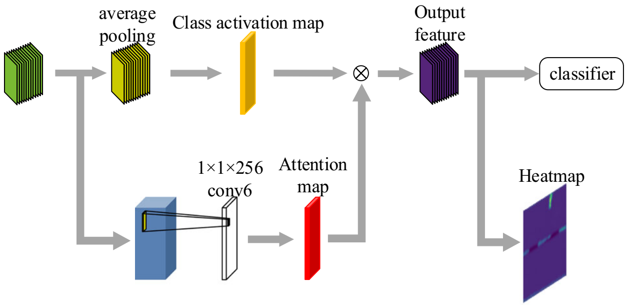


Fig. 3 SA-CAM structure

3 Experiments results analysis and discussion

This section has six parts: dataset's introduction, parameters selection, classsification results, segmentation results, time comparision and analysis. In the experiment, three typical databases with different backgrounds are used. The proposed framework is constructed on an Intel Xeon E5 desktop computer workstation with 8 cores and 64 GB memory. A TITAN-XP graphic card is used to speed up training processing.

3.1 Introduction of image datasets with different backgrounds

In order to evaluate the effectiveness of the proposed method in the context of polymorphic surfaces, three typical data sets with different backgrounds are selected, which are the image dataset with repeated pattern and uniform texture surface, image dataset with homogeneously textured surface, and image dataset with non-homogeneously random texture surface. The three typical databases cover all surface texture types except for the first category [1]. The backgrounds and textures of these datasets are distinct from each other in grayscale, shape, and location, which will help to verify the generalization of the proposed method by following experiments. Next, the characteristics of the data set are described in detail.

3.1.1 DAGM2007 dataset

The first dataset is from DAGM2007 (https://resources.mpiinf.mpg.de/conferences/dagm/2007/prizes.html). It includes six kinds of defects on different texture. Each kind of defect has 1000 non-defect images and 150 defect images. The image size is 512*512 pixels. Labels of all the images are given in the database, but the ground truth for defect areas is given by bounding boxes. Each defect has strong background interference, and the several typical defects are shown in Fig. 4.

The defect types are: linear, planar and irregular. They are different in shape and size under strong background interference. Moreover, the contrasts between the defects and the background are weak. Due to the strong interference of the background and the different shapes of the defect, it brings difficulties to the traditional manual feature extraction and defect segmentation tasks.

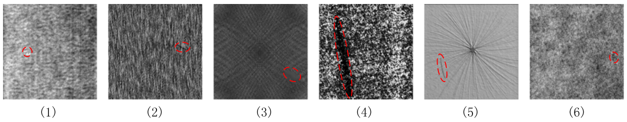


Fig. 4 Examples of DAGM surface defect dataset

3.1.2 Wood knot surface dataset

The second database is the wood defect database [27]. The images also belong to homogeneously textured surface. There are two subsets inside the database. One is used for classification labeled images of defects, which includes different types of wood knots. A total of 438 images from seven types of knot defects are provided in this subset. The other subset of this database is wood board images where the ground truth is provided by bounding boxes. A total of 839 board images are provided including all kinds of defect in first dataset. Some examples are shown in Fig. 5. The images in these two subsets are in different forms and size. At the same time, there are interferences in the edge regions.

3.1.3 Solar cell surface defect

The third data set is a poly-silicon solar cell surface defect dataset from [11]. It belongs to non-homogeneously textured surface. At the same time, the solar cell grid line introduces repeat pattern characteristic to the surface. This cases from a real-world solar cell detecting workshop. The structure is shown in Fig. 6. The surface defects of solar cells in the visible light spectrum range include chipping, broken gates, leaky paste, dirty sheets, scratches, thick lines, and chromatic aberrations. It shows a big difference

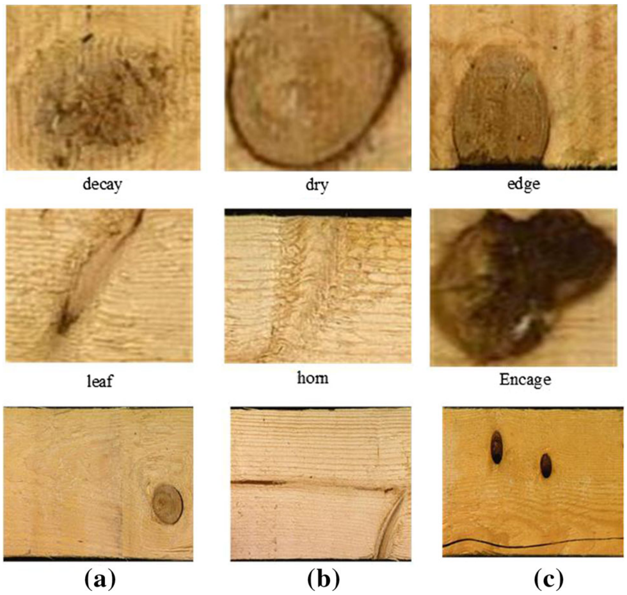


Fig. 5 Examples of wood defect dataset. The first six images show the sub-dataset for classification; the last 3 images are from segmentation sub-dataset

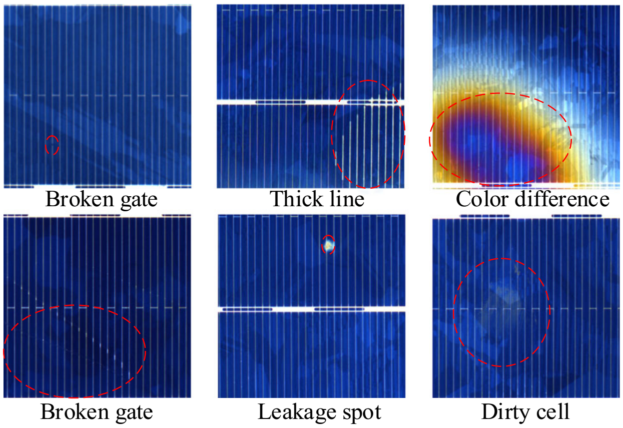


Fig. 6 Examples of solar cell surface defect dataset. The type and location of the defect have been marked in the corresponding position in the figure

between the shape, size, and spectrum characteristics of each defect. The original size of the images is 1828 × 1828 pixels. Broken gate refers to the breakage and loss of the printed finger lines on the surface of the cell. Paste spot is the dripping of the paste when the cell sheets are printed the grid. Dirty cell refers to large dust or dirt on the solar cell. The thick line indicates that the printed weight of the cell sheet is too heavy and the thickness of the gate line is uneven. Scratches are caused by a sharp object passing over the cell. The complex background and random variation of lattice texture and defect character of polycrystalline silicon solar cells bring great challenges for deep learning classification and detection. The dataset obtains 15,330 undefective images and 5915 defective images. The types of defects include broken gates, paste spot, dirty cell, thick lines, scratches, and color differences.

3.2 Parameter selection of convolutional neural network

The model proposed in this paper involves three types of parameters. The first type is image-related parameters, the second type is related parameters in CNN. The third type is related to random forest classifier. The parameters of first type are mainly the following: patch size m and stride size s. According to Ren et al. [19] and Chen et al. [11], the selection of m depends on the domain knowledge for the size of defect. If it is too small, the final heatmap may be inaccurate. On the contrary, the heatmap may be unable to include enough information of the defection. In order to ensure the m and s, three different sets of parameters are prepared and tested separately in DAGM2007. Figure 7 shows the results of three different strides based on fivefold cross-validation.

Table 2 Classification result on DAGM2007

Method	Accuracy
Wavelet representations	95.91
Weibull	97.13
SIFT and ANN	98.24
Tree2vector	96.92
Solar cell CNN	99.26
Solar cell CNN + SVM	97.74
RWSLDC	99.85

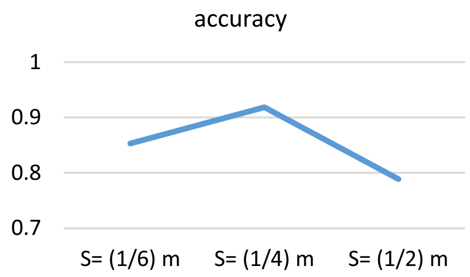


Fig. 7 Results of different strides of and patch

According to Table 2, s is chosen as the (1/4) m in all experiments. The second type of parameters are follows: learning rate λ, training epoch, and dropout ratio. These parameters can significantly affect the features and training results extracted by the CNN. According to [11], the learning rate of the CNN model is selected as $\lambda = 0.0001$, and the epoch of training is 100. The Dropout neuron ratio is 50%. The experiments are showed in Fig. 8.

About the third kind of parameter is related to random forest classifier, the number of decision trees is set to 150, and the maximum depth is set to 200. If the depth and number of trees are further increased, it may affect the algorithm's final detection time. The experiments are shown in Fig. 9.

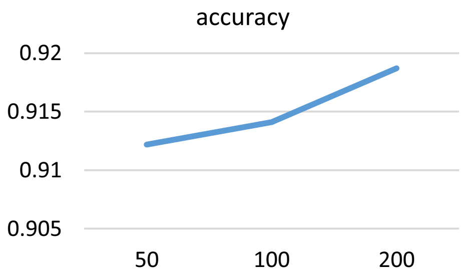


Fig. 8 Results of different epoch

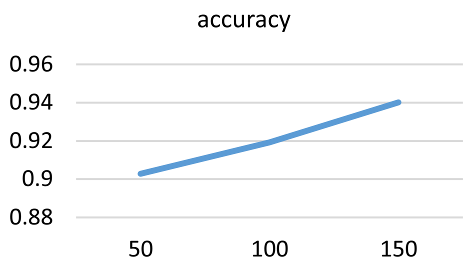


Fig. 9 Results of different strides of and patch

3.3 Classification results and analysis

Based on the above databases, classification and segmentation performance of the presented methods in this paper are evaluated. Each part will be divided into three sections and show the results of experiments and comparative experiments according to dataset. The accuracy reported in this section are all based on fivefold cross-validation unless otherwise stated.

3.3.1 Classification results on DAGM 2007

In this dataset, the comparing experiments includes 12-class CNN, statistical features, SIFT with ANN, and Weibull. The accuracy is shown in Table 2. In this experiment, the comparison results from wavelet representations to SIFT method are from [28], and the tree to vector method is from 6 [29].

From Table 2, it can be seen that: firstly, CNN-based method has a performance increment of at least 1.02% for defect detection problems compared to traditional manually feature-extracting methods because CNN can extract more comprehensive defect features than comparing methods. Secondly, among CNN-based methods, our RWSLDC method increases 0.6% accuracy compared with the original solar cell CNN. That is to say, for different CNN-based methods, even if the features extracted by CNN are identical, random forest classifier shows strong robustness and adaptability. As to SVM classifier, the robustness for multi-class problem is weaker than original method because it is designed for binary classification.

3.3.2 Classification results on wood surface

The method is compared against [19] and the cross-validation methods also are described in the paper. The image feature in [19] is Gabor filters and classifiers include both self-organizing neural network and a feed-forward neural network (FFPNN). The types of defect reported in these methods include encased, leaf, edge, and sound knots. The accuracy is shown in Table 3.

Table 3 Classification accuracy on wood

Method	Accuracy
GLCM + GBC	57.75
GLCM + SVM	60.21
MLBP + MLR	68.64
MLBP + SVM	73.42
MLBP + GBC	75.00
GLCM + MLR	76.32
Gabor + SONN	85.56
Gabor + FuzzySONN	88.34
Gabor + FFPNN	91.17
Tree2vector	90.53
Decaf + MLR	94.29
Solar cell CNN + SVM	97.74
RWSLDC	98.14

From Table 3, comparing to DAGM dataset, due to various defects types and different characteristics, the performance of manually feature-extracting methods on wood dataset is significantly lower. At the same time, CNN-based methods' accuracy is at least 3.12% higher than both general texture features (MLBP and GLCM) and handcrafted classification methods (Gabor filters and FFPNN). At the same time, the proposed RWSLDC method's accuracy reach 98.14%, which is 4% higher than original Decaf and 1% higher than CNN with SVM. That is, though CNN has a strong ability of feature extracting, different classifiers may also lead to difference in performance.

3.3.3 Classification results on solar cell surface defect

This part compares with the work of Chen et al. [11]. They proposed a multi-spectral CNN for the special surface characteristics of solar cells (the solar cell has a blue surface) and used it for defect detection tasks. At the same time, the Gabor and LBP with HOG methods are used as the comparison. In which of these experiments, from LBP to MS-solar cell CNN's results are from [11]. The accuracy is shown in Table 4.

The results show that comparing with the wood dataset, the accuracy increment of the CNN is more obvious than manual extraction feature method, reaching 13%. In this case, CNNs can extract features more effectively. At the same time, among CNN-based methods, the random forest classifier achieved 6% performance improvement than the original CNN. As to the tree2vector method, because different level features are used in the experiments to build the tree-structure and feature vectors, the selected features are still manual features, but the different defects are

Table 4 Classification accuracy on solar cell surface defect

Method	Accuracy
LBP + HOG—SVM	79.26
Gabor + SVM	74.55
Tree2vector	76.90
Solar cell CNN	87.30
MS-solar cell CNN	88.41
Solar cell CNN + SVM	87.26
RWSLDC	93.23

difficult to express with unified features, so the experimental results are not as good as CNN. This shows that the classifier with feature redundancy is more significant for surface defect recognition performance improvement with complex background.

3.3.4 Classification robustness analysis

To explain the classification performance and robustness more intuitively, it is necessary to evaluate the data distribution of the entire dataset. Due to features exacted from CNN are all high-dimensional, the nonlinear dimension reduction method, *t* distributed stochastic neighbor embedding (*t* SNE), is adopted to analyze and visualize the learned features and highlight the useful hidden information in the original module images. Figure 10 illustrates the exacted image features for three different dataset conditions using *t* SNE that can be clearly distinguished. These three experiments evaluate the performance of the proposed solution for different dataset scales. The number of samples for testing the t-SNE effect of the second classification *t* SNE in the DAGM dataset is 1500, and the number of samples for the four types of defect *t* SNE effects of the test surface dataset is 440, and the number of samples for testing the surface defect t-SNE of the solar cell is 1600.

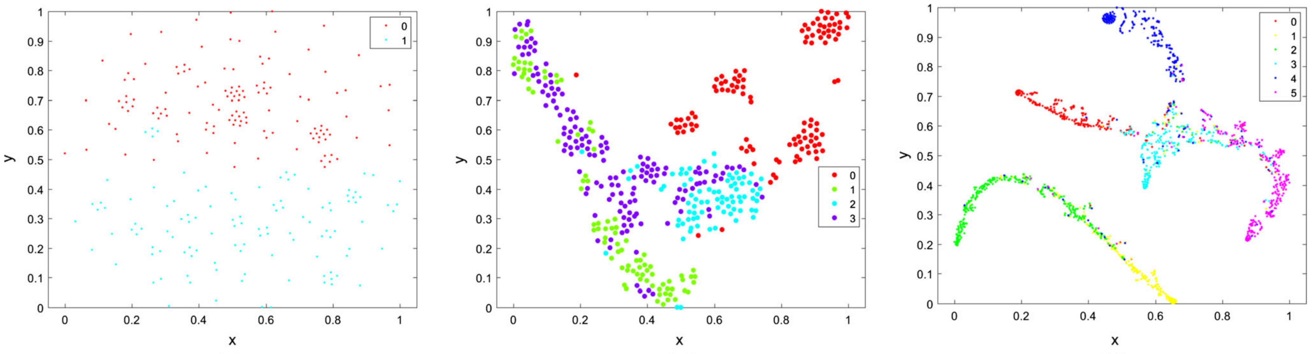
From the t-SNE map, it can be found that as the defect kinds and background texture complexity increase, the classification hyperplane obtained by t-SNE-dimensional reduction becomes more and more complicated. In DAGM2007 dataset, the defect and non-defect images boundary are extremely obvious. However, the extracted features from wood and solar cell surface spread the space and most of them overlap from each other, some of which can be hardly distinguished. Under this environment, the proposed method is still able to get high performance, which verified the classification robustness.

3.4 Segmentation results

The classification results show that the proposed method performs well over multiple datasets. However, there are a few points to note before evaluating the results and presenting the segmentation results. First, since the dataset does not have pixel-level labels, when doing defect segmentation evaluation, we randomly select a certain number of images from the dataset's test set to make a pixel-level ground truth. The used tool is LabelMe, and the link is: https://labelme.csail.mit.edu. Secondly, the performance indicators commonly used in image segmentation are as follows: pixel accuracy (PA), Intersection over Union (IOU), precision (P), and recall(R).

Let there be a total of $k+1$ classes, (from L_0 to L_k, which contains an empty class refers to the background), and the P_{ij} darts belong to class i but is predicted to be the number of pixels of class j. Then P_{ii} represents the number of pixels that are actually predicted, and P_{ij} and P_{ji} blame it as the sum of false positives and false negatives.

Then, the pixel accuracy represents the ratio of the correct pixel to the total pixel, and the result is expressed by



(a) (b) (c)

Fig. 10 t-SNE of learned features using the proposed training model for the three different datasets. **a** The result for DAGM 2007 dataset, **b** the result for wood dataset, **c** the result for solar cell surface dataset

Table 5 Segmentation accuracy on DAGM2007

	1	2	3	4	5	6	average
Decaf							
PA	79.23	46.98	84.16	63.51	68.62	59.73	67.04
IOU	66.48	59.72	70.26	58.64	57.03	48.42	60.09
Precision	70.81	50.29	89.53	62.12	63.75	53.21	64.95
Recall	65.82	43.23	82.15	56.73	56.62	50.9	59.24
F-score	0.6822	0.4649	0.8568	0.5930	0.5997	0.5203	0.6195
Solar cell CNN + CAM							
PA	89.27	65.37	89.6	73.79	89.93	66.8	79.13
IOU	72.94	50.71	74.85	62.02	78.65	50.24	64.90
Precision	72.86	55.83	91.76	66.67	79.87	65.26	72.04
Recall	68.75	46.69	85.21	59.08	76.05	58.82	65.77
F-score	0.7075	0.5082	0.8836	0.7523	0.7791	0.6187	0.6876
SA-CAM							
PA	94.61	66.49	92.89	75.23	92.12	72.81	82.36
IOU	80.33	53.74	83.67	64.26	85.51	55.87	70.56
Precision	90.64	60.19	95.66	67.23	82.14	70.27	77.69
Recall	87.29	53.17	90.12	60.31	78.06	60.34	71.55
F-score	0.8893	0.5646	0.9280	0.6358	0.8004	0.6492	0.7446

$$\mathrm{PA} = \frac{\sum_{i=0}^{k} P_{ii}}{\sum_{i=0}^{k}\sum_{j=0}^{k} P_{ij}}. \tag{7}$$

The mean intersection over union represents the ratio of the intersection and the union of the two sets. In the problem of defect segmentation, the two sets are real values and predicted values. This ratio is also the sum of intersection and intersection, false positive, and negatives. The formula is as follows:

$$\mathrm{mIOU} = \frac{1}{k+1}\sum_{i=0}^{k}\frac{\sum_{i=0}^{k} P_{ii}}{\sum_{j=0}^{k} P_{ij} + \sum_{j=0}^{k} P_{ji} - P_{ii}}. \tag{8}$$

Precision measures the exactness or fidelity of detection and segmentation and is calculated in Eq. (9). Recall describes the completeness of detection and segmentation and is defined in Eq. (10). F-measure combines precision and recall and is computed in Eq. (11). Table 4 shows the precision, recall, and F-measure for the solar cell CNN. (TP represents a true positive, that is, pixels labeled as defective are correctly detected; FP indicates false positives, that is, pixels labeled as good are erroneously detected as defective; FN means false negative, that is, pixels labeled as defective are erroneously detected as non-defective; TN represents a true negative, that is, pixels labeled as non-defect are correctly detected as non-defect)

$$\mathrm{Precision} = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FP}} \tag{9}$$

$$\mathrm{recall} = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FN}} \tag{10}$$

$$\text{F-score} = \frac{2 \times \mathrm{precision} \times \mathrm{recall}}{\mathrm{precision} + \mathrm{recall}} \tag{11}$$

In order to accurately detect defects, the defect detection task requires high IOU and PA. At the same time, maximize the precision while ensuring the recall. Using these four indicators, the results of the segmentation experiment on the three databases are given below.

3.4.1 Segmentation results on DAGM 2007

This dataset has ground truth given by bound boxes. According to these ground truths, the label masks are made by LabelMe. The average index is shown in Table 5.

Table 5 shows that the proposed method achieves a performance improvement of about 5% on the PA and IOU indicators compared to the CAM method in the defect segmentation task. At the same time, Ren's method is also 6% lower than the CAM-based approaches. Some of the result images and comparative experimental results in this dataset are shown in Fig. 11.

From Fig. 8, it can be seen that the proposed method performs well on the database. On the contrast, original CAM interferences segmentation tasks on different levels from the first to fourth types of surface in this dataset. In the second type of defect, the CAM method fails to detect the region where the defect is located. Among the third and fourth types of defects. The CAM method marks a large background area as a defect, but at the same time, the proposed method can suppress the background, making the detection area more accurate finally leading to a better segmentation effect. At the same time, though Ren's method can detect the approximate location of each type of defect, it is difficult to accurately segment the defects.

3.4.2 Segmentation results on wood surface knot

This experiment compares to [30] and comparison experiments in the paper [31, 32], including two manual feature extraction methods and three CNN-based deep-learning method. Like the DAGM2007 dataset, this datasets ground truth is also given by bounding box. The pixel-level ground truth is still annotated manually by LabelMe. Because of the overmuch kinds of defects in the dataset for segmentation, the experimental results will be given as the average value, and the segmentation results of some images of defect will be given. The average PA and IOU of each methods are shown in Table 6.

From Table 6, it can be seen that deep learning methods achieve extraordinary increament (nearly 30%) comparing to the manual features. On the other hand, there are also great differences on IOU, precision, and recall between various deep learning segmentation methods. Some images also are shown in Fig. 12. For these outputs, the white area shows the area of defection.

It is shown that manual feature extracting method is hard to segment defect area on wood surface precisely. For the image (a)–(d), five baseline methods appearance missed or detect of some defect regions incorrectly. Among CNN-based method, in (a) and (d), the proposed method is able to detect the defection area accurately. In (c), the method can extract the area which DCGAN method unable to inspect. Comparing to the original CAM, the proposed method suppresses the high responsive background part of the CAM method while improving the response of the defective portion.

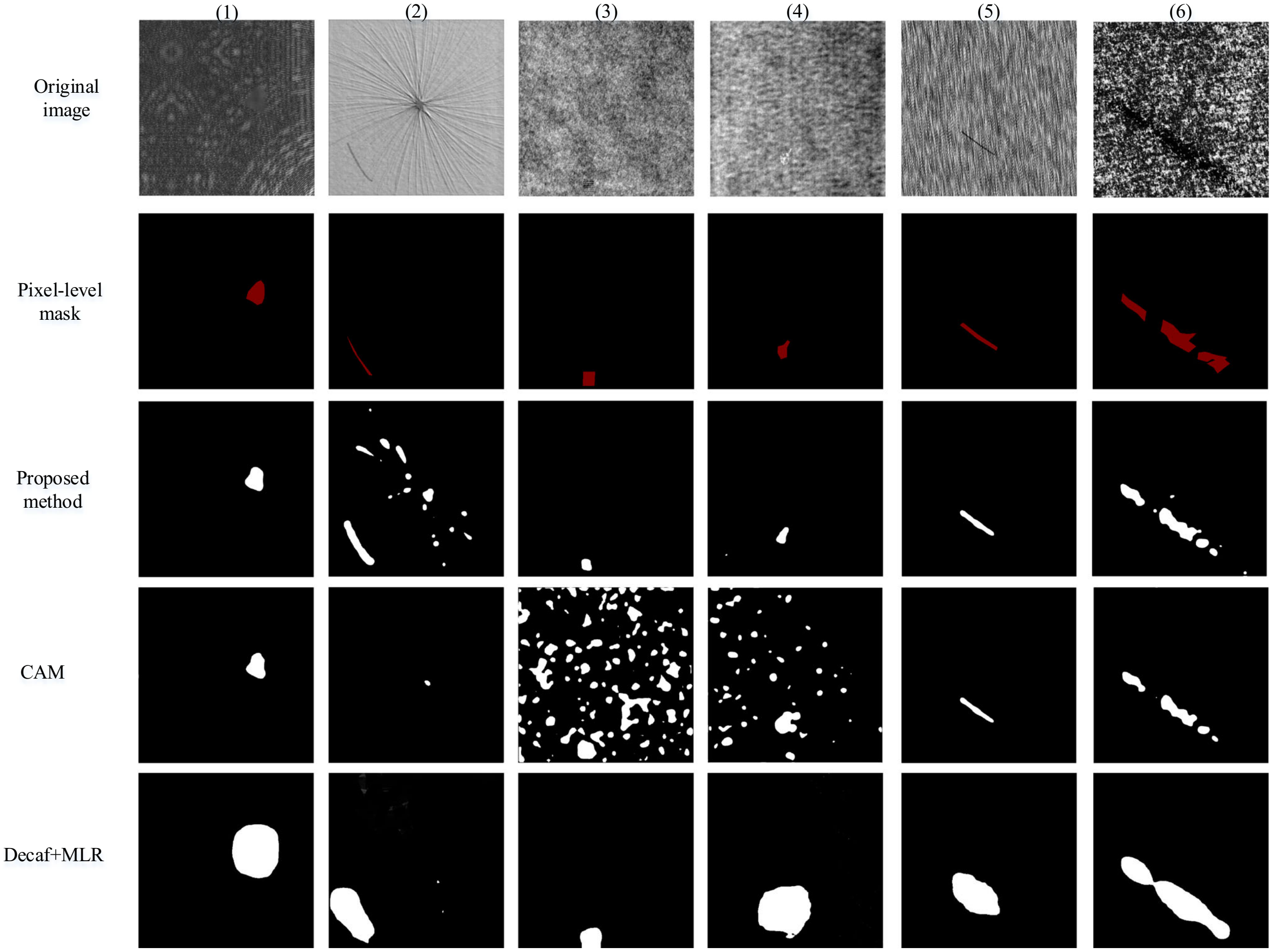


Fig. 11 Result images for proposed and comparative experimental result on DAGM2007

Table 6 Segmentation result on wood surface dataset

Method	PA	IOU	Precision	Recall	F-score
Covariance matrices	41.51	25.02	26.12	15.31	0.1930
Boolean map	47.80	33.89	23.04	17.96	0.2019
Decaf	70.2	57.16	67.27	60.04	0.6345
CAM	73.85	58.63	68.27	52.23	0.5918
DCGAN	79.85	63.92	73.23	63.21	0.6785
SA-CAM	79.54	64.27	75.97	61.21	0.6780

3.4.3 Segmentation results on solar cell surface defect

There is little study on a variety of solar surface defects segmentation, so the benchmark method for this experiment are only CAM and decaf. The experimental results are given in Table 7, including several typical types of defects. Some images also are shown in Fig. 13.

From Table 7 and Fig. 13, it can be seen that solar cell texture surface has a very strong impact on the final segmentation result, resulting in poor results in the comparative experiments. The CAM-based method makes 15% increment comparing to [19] at the three defects except the dirty cell. The proposed methods' effect is still about 4% higher than original CAM's IOU and PA. For dirty cell defection, since the contrast of the defect is too weak

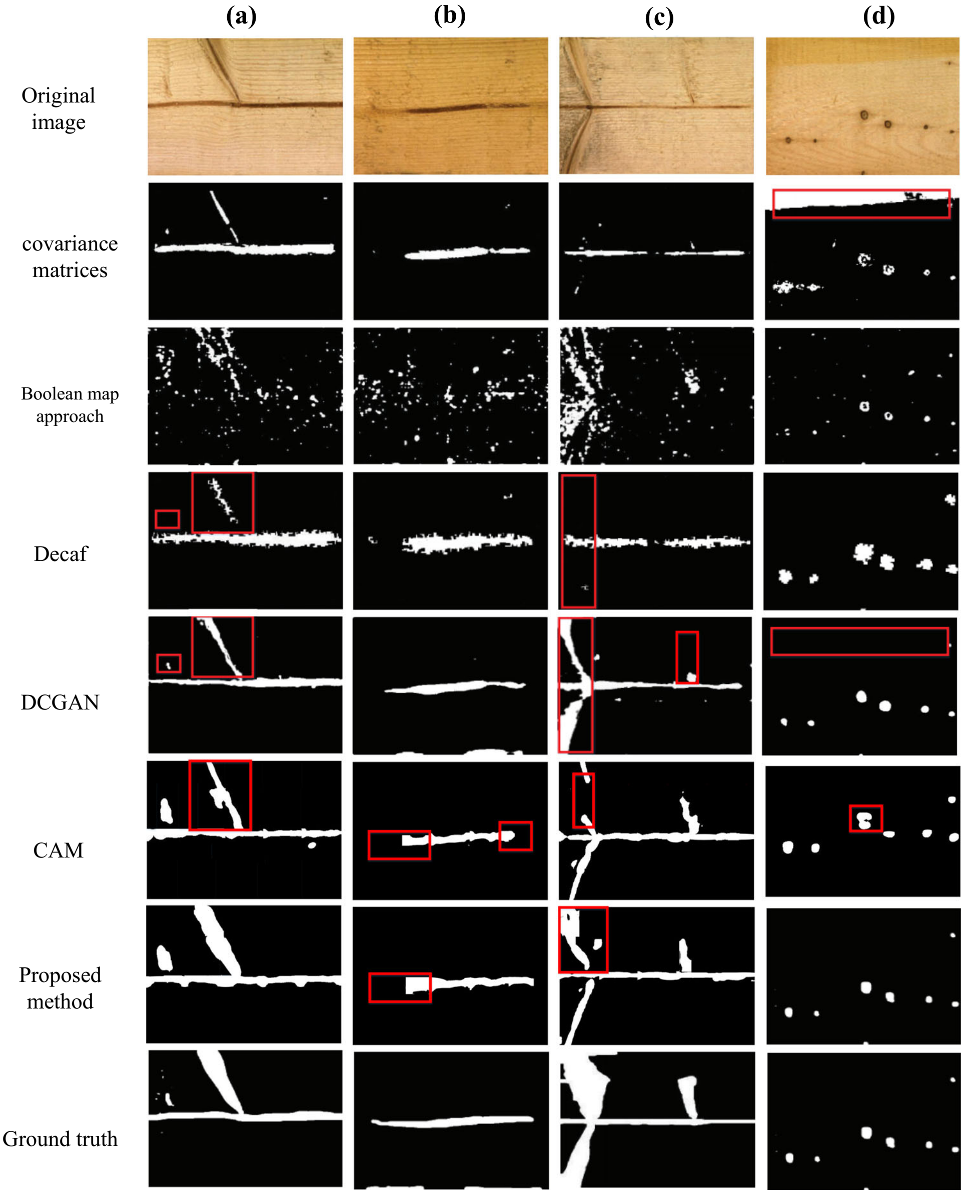


Fig. 12 Result images for proposed and comparative experimental result on wood knot dataset

Table 7 Result images for proposed and comparative experimental results on solar cell surface defect dataset

Index	Paste spot	Dirty cell	Thick lines	Color difference	Average
Decaf + MLR					
PA	52.42	29.45	56.51	43.49	45.48
IOU	40.07	14.25	41.50	31.45	31.82
Precision	60.13	34.03	63.97	42.05	50.05
Recall	49.02	16.76	56.53	35.07	39.34
F-score	0.5401	0.2246	0.6002	0.3824	0.4368
CAM					
PA	75.05	31.86	77.34	59.63	60.97
IOU	61.79	18.04	62.76	49.89	48.12
Precision	76.24	29.12	65.73	52.71	55.95
Recall	64.13	13.24	57.06	40.95	43.85
F-score	0.6966	0.1820	0.6109	0.4609	0.4876
SA-CAM					
PA	82.68	36.13	83.39	65.02	66.805
IOU	65.52	23.07	70.24	55.63	53.62
Precision	80.23	30.61	70.48	55.91	59.30
Recall	68.06	14.75	60.22	47.86	47.72
F-score	0.7362	0.1991	0.6495	0.5157	0.5252

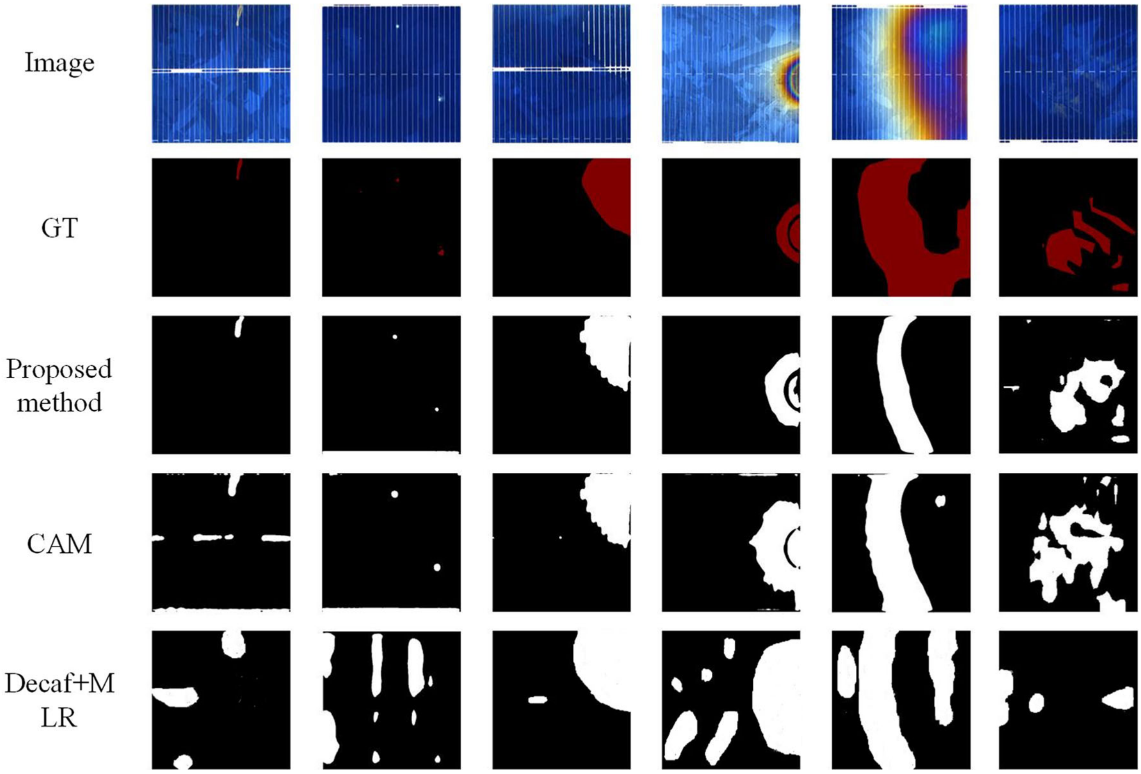


Fig. 13 Result images for proposed and comparative experiments on solar cell surface

compared with the background, the segmentation effect of such defects is the worst among the four types of defects. At the same time, compared with the previous two datasets, the increment of proposed method is the most significant comparing to the comparative experiment. This also verifies that the proposed methods robustness and adaptability to complex surfaces.

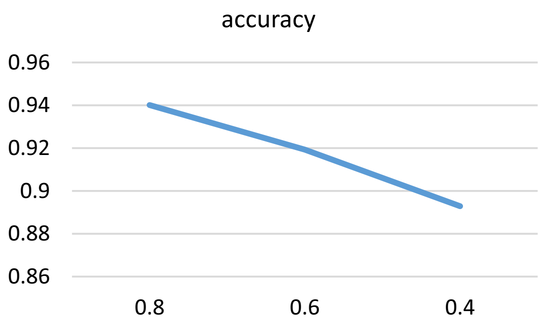


Fig. 14 Results for three different training ratios

3.5 Stability discussion

To verify the stability of proposed method, the ratio of the training and test sets is 8:2, 6:4, and 4:6 respectively. The results are shown in Fig. 14. This experiment is conducted to demonstrate that multi-spectral solar cell convolutional neural networks are still effective when the dataset is still a small percentage of overall production data. Figure 14 shows the results of three experiments.

From Fig. 14, it can be obtained that as the ratio of the training and test sets increases, the accuracy of proposed method increases slightly. When the ratio of training set is 0.4, the classification accuracy will reduce by about five percentage. The experimental results illustrate the stability of unknown defect samples in some extent.

3.6 Time comparison

The time for each method of training and testing is given in Table 8. It can be seen that the SVM classifier requires more training time than the random forest in the multi-classification task, and the detection speed is slower than the random forest classifier. It should be mentioned that the deep learning model can use batch normalize to input batch of images same time, while the traditional machine vision method needs to read and extract features cyclically, so the total classifying time per hundred images is quite different.

3.7 Concluding discussion

From the experimental results, the feature extracted from attention-based CNN outperforms handcrafted features for all kinds of texture. At the same time, the random forest classifier and attention module further enhance the effect of classification. In addition, the SA-CAM also performs well in weakly supervised defect segmentation.

3.7.1 Classification

The experimental results show that CNN can be used as a general feature extractor for different surfaces and defects compared to the traditional manually selecting features' method. In addition, by replacing the fully connected layer and replacing it with different classifiers and introducing attention mechanisms, it can achieve better classification results than traditional CNN and has better generalization performance than traditional methods.

3.7.2 Segmentation

Through the segmentation experiment and performance evaluation of several typical image datasets, it can be found that SA-CAM has the following advantages: (1) The model's robustness is strong. In the DAGM dataset, the model accurately inspects defect locations and segmented defect regions in six kinds of distinct repeated pattern texture surfaces and defects. Defect inspection and segmentation can also be achieved for dozens of different defects randomly distributed on the homogeneously textured wood surface and non-homogeneous textured like solar cell surface with grid lines and crystal lattice. (2) Defect segmentation of weak-supervised learning and reduction in the dependence on the pixel-labeled samples are achieved. The proposed method only requires global image-level annotation to achieve accurate multi-class surface defect recognition.

4 Conclusion

This paper proposes a robust weakly supervised learning of deep CNN framework (RWSLDC) for automatic surface inspection. This framework is able to solve the problems of automatic surface inspection by using global image-level labels during training. For the classification task, the proposed framework achieves a better classification effect than

Table 8 Classifying time of some methods

	Training time (s)	Detecting time (100 images)
RWSLDC	4987	4.05
CNN with SVM	5103	4.23
CNN	4869	3.66
LBP + HOG-SVM	9785	42.20
Gabor-SVM	9670	35.70

the traditional manual extraction feature and the basic CNN [11] through the attention mechanism and random forest classifier. The accuracy increased by 0.6–25.5%. In segmentation task, the SA-CAM under framework increases defect segmentation's pixel accuracy and intersection ratio by 5.49–7.07%. At the same time, the proposed method does not require precise pixel-level marking, thus reducing the cost of expensive dataset marking, which reduce the costly manual annotation.

In the future, we will focus on two directions of research. One direction involves the speeding up the heatmap generation process to enable real-time defect localization. The second direction is looking for more efficiency attention module. In addition, considering the difference between defect and background, the defect sample can be regarded as a tree-structured data, so the method of the tree2vector class is worth further research.

Acknowledgements This work is supported in part by National Natural Science Foundation (NNSF) of China under Grant 61873315, Natural Science Foundation of Hebei Province under Grant F2018202078, Science and Technology Program of Hebei Province under Grant 17211804D, Hebei Province Outstanding Youth Science Foundation F2017202062 and Young Talents Project in Hebei Province under Grant 210003.

Compliance with ethical standards

Conflict of interest We declare that we have no financial and personal relationships with other people or organizations that can inappropriately influence our work, there is no professional or other personal interest of any nature or kind in any product, service, and company that could be construed as influencing the position presented in or the review of the manuscript entitled.

References

1. Xie X (2008) Review of recent advances in surface defect detection using texture analysis techniques. ELCVIA: Electron Lett Comput Vis Image Anal 7(3):1
2. Luo Q, Sun Y, Li P, Simpson O, Tian L, He Y (2018) Generalized completed local binary patterns for time-efficient steel surface defect classification. IEEE Trans Instrum Meas 68(3):667
3. Binyi S et al (2019) Classification of manufacturing defects in multicrystalline solar cells with novel feature descriptor. IEEE Trans Instrum Meas 68(12):4675–4688
4. Yapi D, Allili MS, Baaziz N (2017) Automatic fabric defect detection using learning-based local textural distributions in the contourlet domain. IEEE Trans Autom Sci Eng 15(3):1014
5. Wang H, Qi H, Wang XF (2013) A new Gabor based approach for wood recognition. Neurocomputing 116:192
6. Zhang Z, Zou Y, Gan C (2018) Textual sentiment analysis via three different attention convolutional neural networks and cross-modality consistent regression. Neurocomputing 275:1407
7. Xie L, Huang R, Gu N, Cao Z (2014) A novel defect detection and identification method in optical inspection. Neural Comput Appl 24(7):1953
8. Liu W, Wang Z, Liu X, Zeng N, Liu Y, Alsaadi FE (2017) Survey of deep neural network architectures and their applications. Neurocomputing 234:11
9. Mirjalili SM, Mirjalili SZ (2017) Single-objective optimization framework for designing photonic crystal filters. Neural Comput Appl 28(6):1463
10. Jung S, Tsai Y, Chiu W, Hu J, Sun C (2018) Defect detection on randomly textured surfaces by convolutional neural networks. In: 2018 IEEE/ASME international conference on advanced intelligent mechatronics (AIM) (IEEE, 2018), pp 1456–1461
11. Chen H, Pang Y, Hu Q, Liu K (2020) Solar cell surface defect inspection based on multispectral convolutional neural network. J Intell Manuf 31:453–468
12. Zhou S, Chen Y, Zhang D, Xie J, Zhou Y (2017) Classification of surface defects on steel sheet using convolutional neural networks. Mater Technol 51(1):123
13. Tang Y (2013) Deep learning using linear support vector machines. arXiv:1306.0239
14. Merentitis A, Debes C (2015) Automatic fusion and classification using random forests and features extracted with deep learning. In: 2015 IEEE international geoscience and remote sensing symposium (IGARSS) (IEEE, 2015), pp 2943–2946
15. Zhang H, Zhang L, Li P, Gu D (2018) Yarn-dyed fabric defect detection with yolov2 based on deep convolution neural networks. In: 2018 IEEE 7th data driven control and learning systems conference (DDCLS) (IEEE, 2018), pp 170–174
16. Singh J, Shekhar S (2018) Road damage detection and classification in smartphone captured images using mask r-cnn. arXiv:1811.04535
17. Yuille AL, Liu C (2018) Deep nets: What have they ever done for vision? arXiv:1805.04025
18. Zhou B, Khosla A, Lapedriza A, Oliva A, Torralba A (2016) Learning deep features for discriminative localization. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 2921–2929
19. Ren R, Hung T, Tan KC (2017) Generic deep-learning-based approach for automated surface inspection. IEEE Trans Cybern 48(3):929
20. Lin H, Li B, Wang X, Shu Y, Niu S (2019) Automated defect inspection of led chip using deep convolutional neural network. J Intell Manuf 30(6):2525
21. Li W, Leonardis A, Fritz M (2017) Visual stability prediction and its application to manipulation. In: 2017 AAAI Spring symposium series
22. Jaderberg M, Simonyan K, Zisserman A et al (2015) Spatial transformer networks. In: Advances in neural information processing systems, pp 2017–2025
23. Ji Y, Zhang H, Wu QMJ (2018) Salient object detection via multi-scale attention CNN. Neurocomputing 322:130–140
24. Breiman L (2001) Random forests. Mach Learn 45(1):5
25. Kairanbay M, See J, Wong LK, Hii YL (2017) Filling the gaps: reducing the complexity of networks for multi-attribute image aesthetic prediction. In: 2017 IEEE international conference on image processing (ICIP) (IEEE, 2017), pp 3051–3055
26. Otsu N (1979) A threshold selection method from gray-level histogram. IEEE Trans Syst Man Cybern 9(1):62
27. Silven O, Niskanen M, Kauppinen H (2003) Wood inspection with non-supervised clustering. Mach Vis Appl 13(5–6):275
28. Wang T, Chen Y, Qiao M, Snoussi H (2018) A fast and robust convolutional neural network-based defect detection model in product quality control. Int J Adv Manuf Technol 94(9–12):3465
29. Zhang H et al (2018) Tree2Vector: learning a vectorial representation for tree-structured data. IEEE Trans Neural Netw Learn Syst 99:1–15
30. Zhai W, Zhu J, Cao Y, Wang Z (2018) A generative adversarial network-based framework for unsupervised visual surface

inspection. In: 2018 IEEE international conference on acoustics, speech and signal processing (ICASSP) (IEEE, 2018), pp 1283–1287

31. Zhang J, Sclaroff S (2015) Exploiting surroundedness for saliency detection: a Boolean map approach. IEEE Trans Pattern Anal Mach Intell 38(5):889

32. Donoser M, Bischof H (2008) Using covariance matrices for unsupervised texture segmentation. In: 2008 19th international conference on pattern recognition (IEEE, 2008), pp 1–4

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.