Mechanical Systems and Signal Processing 175 (2022) 109109

Contents lists available at ScienceDirect

Mechanical Systems and Signal Processing

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

Fusion of multi-light source illuminated images for effective defect inspection on highly reflective surfaces

Guizhong Fu [a,b,c,1], Shukai Jia [a,c,1], Wenbin Zhu [a,c], Jiangxin Yang [a,c], Yanlong Cao [a,c], Michael Ying Yang [d], Yanpeng Cao [a,c,*]

[a] State Key Laboratory of Fluid Power and Mechatronic Systems, School of Mechanical Engineering, Zhejiang University, Hangzhou 310027, China
[b] School of Mechanical Engineering, Suzhou University of Science and Technology, Suzhou, China
[c] Key Laboratory of Advanced Manufacturing Technology of Zhejiang Province, School of Mechanical Engineering, Zhejiang University, Hangzhou, China
[d] Scene Understanding Group, University of Twente, Hengelosestraat 99, 7514 AE Enschede, The Netherlands

ARTICLE INFO

Communicated by Y. Lei

Keywords:
Surface inspection
Defect classification
Convolutional neural network
Image fusion
Multi-light source illumination

ABSTRACT

It is observed that a human inspector can obtain better visual observations of surface defects via changing the lighting/viewing directions from time to time. Accordingly, we first build a multi-light source illumination/acquisition system to capture images of workpieces under individual lighting directions and then propose a multi-stream CNN model to process multi-light source illuminated images for high-accuracy surface defect classification on highly reflective metal. Moreover, we present two effective techniques including individual stream deep supervision and channel attention (CA) based feature re-calibration to generate and select the most discriminative features on multi-light source illuminated images for the subsequent defect classification task. Comparative evaluation results demonstrate that our proposed method is capable of generating more accurate recognition results via the fusion of complementary features extracted on images illuminated by multi-light sources. Furthermore, our proposed light-weight CNN model can process more than 20 input frames per second on a single NVIDIA Quadro P6000 GPU (24G RAM) and is faster than a human inspector. Source codes and the newly constructed multi-light source illuminated dataset will be accessible to the public.

1. Introduction

Effective surface defect inspection (detection/classification) is an important processing step in many industrial production tasks for the quality control of end products [1]. Professional human inspectors are typically trained to perform visual quality of end products which is a highly subjective, labor-intensive, and time-consuming task. As the result, developing accurate, real-time and fully automatic inspection solutions has received significant attention in recent years [2].

Numerous machine vision-based surface inspection methods have been proposed as an effective way for non-contact, fast, and automatic surface defect detection/classification tasks [3–6]. Many existing approaches deployed Fourier transform [7], wavelet filters [8,9] or Gabor filters [10,11] to extract representative feature maps of defects, and then trained support vector machines [8,12,13] or neural networks [14,15] to predict the existence of defects (i.e., defect detection) and to determine

* Corresponding author.
E-mail address: caoyp@zju.edu.cn (Y. Cao).
[1] These authors contributed equally to this work.

https://doi.org/10.1016/j.ymssp.2022.109109
Received 31 January 2021; Received in revised form 15 March 2022; Accepted 27 March 2022
Available online 19 April 2022

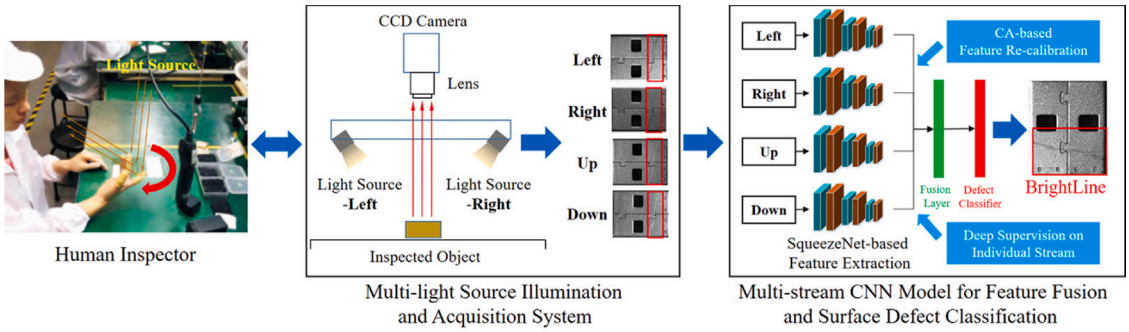

Fig. 1. The built multi-light source illumination and acquisition system and the proposed multi-stream CNN model for extraction and fusion of multi-light source illuminated features for surface defect classification.

their corresponding categories (i.e., defect detection). However, the above-mentioned methods built on the hand-crafted features cannot generate accurate detection/classification results when surface defects present "inter-class" similarity and "intra-class" diversity [16,17].

Due to the powerful learning capacity, Convolutional Neural Network (CNN) has been utilized as a prevalent tool to generate highly distinctive features in various computer vision tasks including object classification [18–21] and target detection [22,23] and achieved significant performance enhancement. VGG model deploys a number of 3×3 filters in different convolutional stages to extract multi-scale features which are further utilized for various computer vision tasks [24]. ResNet is a residual learning framework which adds skip connections between shallower and deeper layers to overcome the vanishing gradient problem occurred when training deep CNN models [25]. To achieve higher computational efficiency, a number of compact CNN architectures such as SqueezeNet [26] and YOLO [27] have been developed for accurate target recognition using significantly fewer parameters.

In recent decades, many researchers attempted to build CNN-based models for surface defects defection/classification [14,16,28–30]. The underlying principle of these methods is to directly learn highly discriminative features from training samples of various defects for subsequent defection/classification tasks. A compact CNN model is built to compute the symmetric surround saliency map for recognition of seven types of defects on steel surface [28]. However, the performance of this model is not satisfactory since it is built from scratch and its parameters are randomly initialized. In [17], a deep convolutional activation feature extractor (Decaf model [31]) is integrated with a multinomial logistic regression classifier to perform generic and high accuracy surface defect classification task. Note the Decaf architecture [31] pre-trained in ImageNet is directly utilized without parameter fine-tuning thus the extracted features are not optimal for the subsequent defect inspection tasks. A CNN model is proposed for accurate and fast classification of surface defects under severe illumination and noise disturbances by emphasizing the parameter tuning of shallower convolutional layers and integrating multiple-scale features [32]. A number of convolutional layers with multiple receptive fields are stacked in a CNN model for accurate classification of large- and small-size defects in cluttered background [33]. An effective triplet-graph reasoning network is proposed to improve the poor generalization performance when the model is trained using insufficient samples [34]. To overcome the difficulty of pixel-level image labeling, a novel image-level weakly supervised segmentation formulation is proposed for no-service rail surface defects [35]. Note the above-mentioned surface defect inspection methods typically utilize a single light source to illuminate the target workpieces for image acquisition. They assume that the surface defects are visually observable local anomalies on single input images which cannot be well satisfied in many practical industrial inspection tasks particularly for products made of highly reflective materials [36–38].

In this paper, we present an effective methodology for accurate and fast defect inspection on highly reflective surfaces based on the fusion of images illuminated by multi-light sources. Our intuition is that a human inspector will change the lighting/viewing directions from time to time, obtaining better visual observations of surface defects for more accurate detection/classification as illustrated in Fig. 1. Accordingly, we build a multi-light source illumination system to capture images of workpieces under four different lighting directions (Left, Right, Up, and Down) and then propose a multi-stream CNN model to extract and integrate features on multi-light source illuminated images for surface defect classification, as illustrated in Fig. 1. To achieve high-accuracy defect inspection on highly reflective surfaces, we adopt the pre-trained SqueezeNet architecture to build a multi-stream model and incorporate deep supervision to enhance the discriminative ability of features extracted on individual streams. Moreover, we utilize the channel attention (CA) mechanism to re-calibrate the feature responses towards the more representative channels and thus select the most discriminative features under multiple illumination directions for the subsequent defect classification task. Evaluated on a newly constructed dataset containing different types of defects on highly reflective stainless steel, our proposed defect classification model achieves a significant performance gain through the effective fusion of complementary features extracted on images illuminated by multi-light sources. Our proposed method can generate more accurate results compared with some best-performing CNN-based defect classification models [16,17,28,32,39]. Furthermore, the proposed multi-stream CNN model can process over 20 input frames on a single NVIDIA Quadro P6000 GPU (24G RAM) and can satisfy the demand of on-line surface defect detection tasks. As illustrated in Fig. 2, our proposed method could be easily deployed in a realistic industrial scene to facilitate automatic product manufacturing and inspection. To sum up, the contributions of this work include:

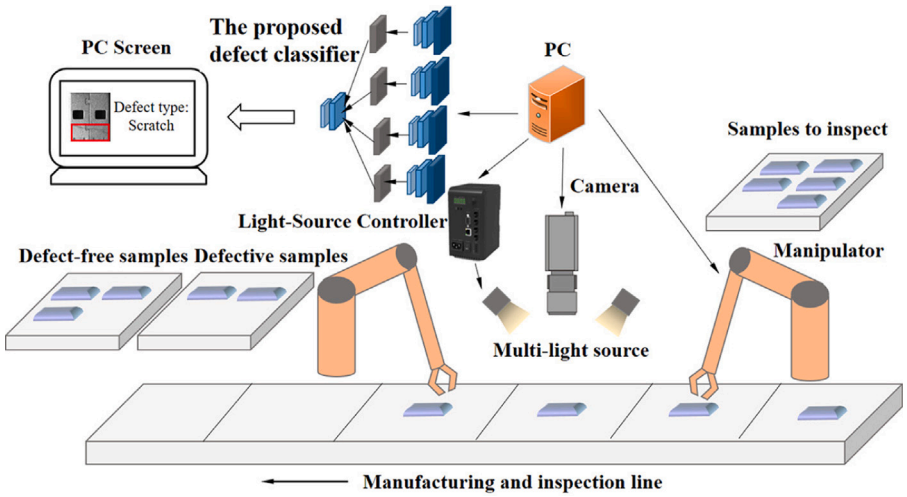

Fig. 2. The schematic illustration of the deployment of the proposed method in a realistic industrial manufacturing and inspection process.

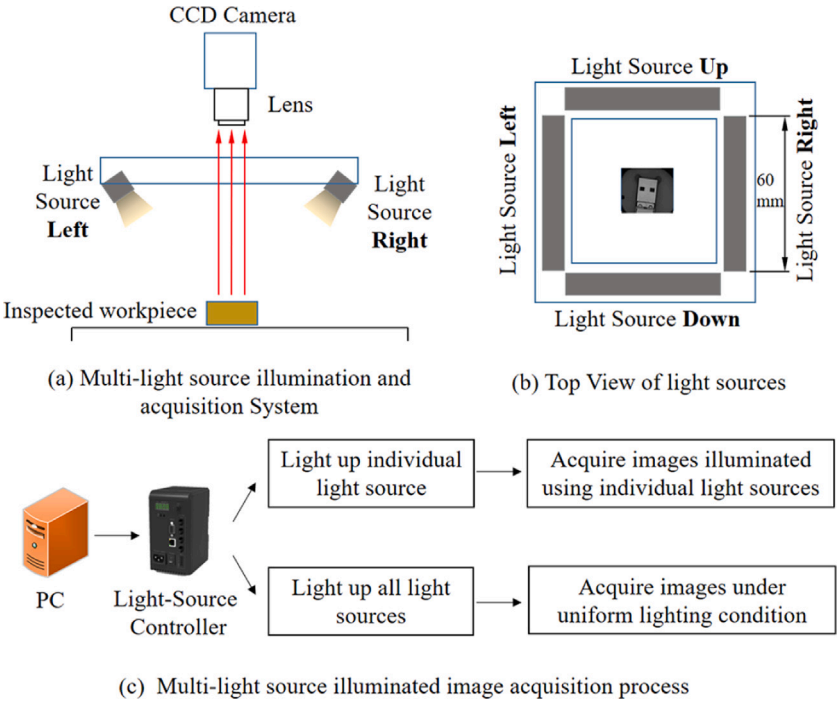

Fig. 3. The setup of the multi-light source illumination/acquisition system.

(1) We build a multi-light source illumination/acquisition system which can capture images of workpieces under individual lighting directions (Left, Right, Up, and Down) and construct a new multi-light source illuminated dataset containing images of different types of surface defects on highly reflective stainless steel.

(2) We incorporate deep supervision in our proposed multi-stream CNN model to reinforce the discriminative ability of individual feature extraction streams and utilize the channel attention (CA) mechanism to select the most discriminative features under multiple illumination directions for the subsequent defect classification task.

(3) Our method effectively extract and integrate complementary features on images captured under different lighting directions and thus achieves significantly improved recognition performance compared with some best-performing defect classification models [17,28,32]. In addition, this light-weight CNN model (3.5 MB) can process more than 20 input frames per second using a single NVIDIA Quadro P6000 GPU to facilitate online inspection tasks.

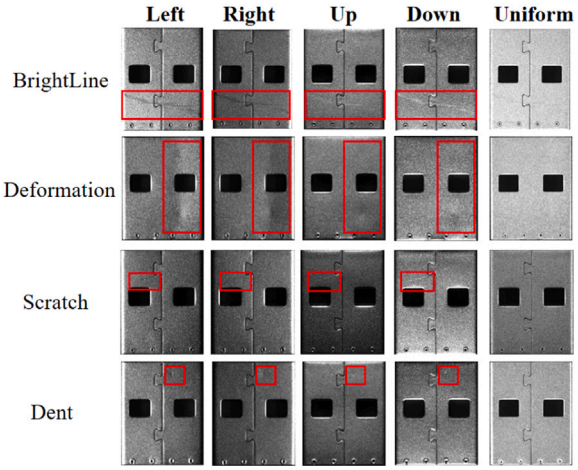

Fig. 4. Sample images of BrightLine, Deformation, Scratch, and Dent surface defects illuminated using individual light sources.

2. Multi-light source illumination/acquisition system and dataset

It is important to set up appropriate lighting configurations and capture images with good observations of target objects for high-quality visual inspection tasks (e.g. surface defect detection and classification). It is noted that human inspectors can better identify various types of surface defects with different locations and orientations by changing the lighting/viewing directions from time to time. Based on this intuition, we build a multi-light source illumination/acquisition system to capture images of workpieces under different lighting directions.

Fig. 3(a)–(b) show the hardware configuration of multi-light source illumination/acquisition system. A monochrome industrial camera (2448 × 2048 resolution) and a bi-telecentric lens (260 mm working distance) are deployed to capture images of target workpieces. Four 24 V LED strip lights (60 mm in length) are installed in a squared aluminum frame to generate light stimulation from four individual directions (Left, Right, Up and Down) to better illuminate the insignificant surface defects. The incident angles of strip light sources are all set to 45°.

Using the built multi-light source illumination/acquisition system, we capture images of workpieces made of highly reflective stainless steel (USB connectors). As shown in Fig. 3(c), the entire image acquisition process includes two major steps. First, we send commands to the light source controller to turn on four light sources one by one and capture images illuminated from four individual directions (Left, Right, Up, and Down). Then, we turn on all the light sources at the same time to capture another image under the uniform lighting condition. Therefore, five images are captured for each workpiece including four single-light source illuminated images and a uniformly illuminated one. It is observed that the surface defects present very different/complementary characteristics under individual lighting conditions as shown in Fig. 4. Moreover, some defects can be well observed in multi-light source illuminated images but become indistinct in the uniformly illuminated image.

In total, we captured multi-light source illuminated images (four single-light source illuminated images and one uniformly illuminated image) of 700 samples (USB connectors) with four typical surface defects including BrightLine, Deformation, Scratch, and Dent. The collected images are split into the training and testing datasets with an aspect ratio of 1:4. More specifically, the training dataset consists of images of 560 USB connectors (2,800 images in total) and the testing dataset contains images of a new batch of 240 USB connectors (1,200 images in total). Note we use different batches of workpieces to capture training and testing sample images, therefore defects of the same class present different visual characteristics in the training and testing dataset.

It is noted that the captured images typically contain redundant background regions which incur undesirable computational overhead in industrial inspection tasks. We adopted the technique presented in [32] to define a rectangle bounding box on the input image, which can be used to appropriately cover the target workpiece and remove the redundant background, as illustrated in Fig. 5(a). Since the position of the workpiece remains unchanged during capturing multi-light source illuminated images, we make use of the image under uniform lighting to calculate the coordinates of the ROI which are then mapped to other multi-light source illuminated images to define their corresponding ROI areas. In the extracted ROIs, defective regions are manually defined/labeled by human inspectors (viewing five multi-light source illuminated images) and then uniformly divided into a number of image patches (200 × 200) as illustrated in Fig. 5(b). In addition, we include 3,600 and 900 image patches without defects (Normal) to the training and testing datasets, respectively. In Table 1, we summarize the number of manually labeled image patches of Normal, BrightLine, Deformation, Scratch, and Dent classes in the newly constructed multi-light source illuminated dataset. Some sample images of defectives and normal surfaces illuminated using individual light sources are shown in the left of Fig. 5(b).

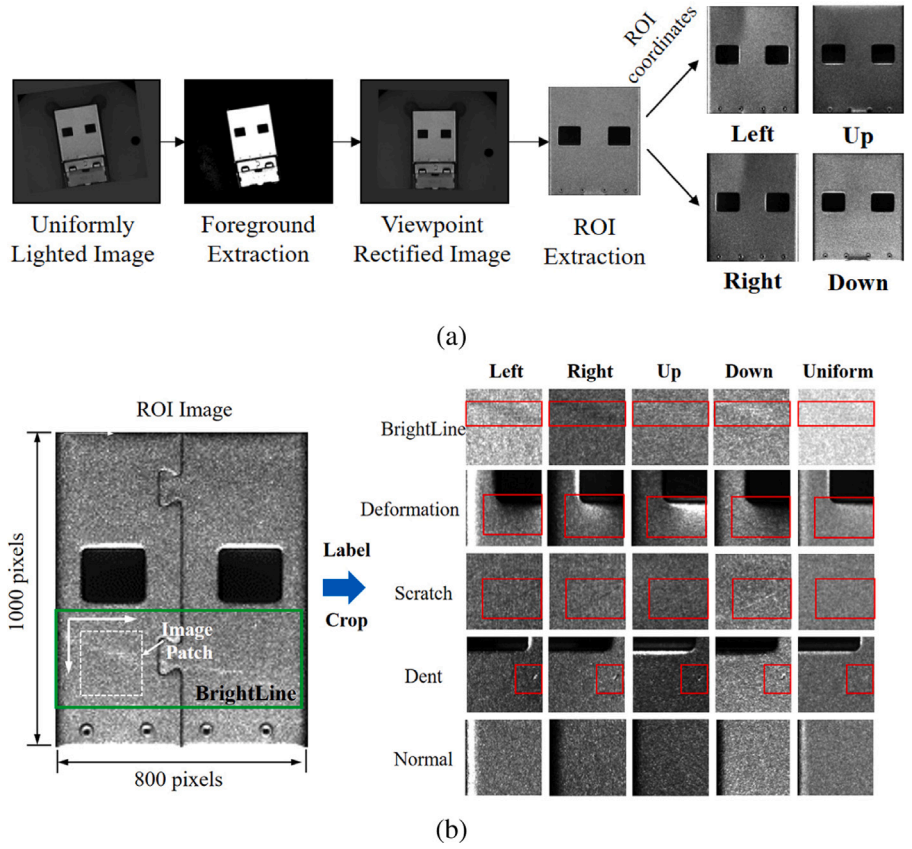

Fig. 5. The process of extracting ROIs and generating a number of labeled image patches on images illuminated using different light sources. (a) ROIs extraction and (b) Image patch cropping.

Table 1
The number of labeled image patches of Normal, BrightLine, Deformation, Scratch, and Dent in the newly constructed multi-light source illuminated dataset.

Defect type	Training dataset	Testing dataset
BrightLine	2072	573
Deformation	2511	541
Dent	1282	276
Scratch	2265	425
Normal	3600	900
Sum	11730	2715

3. Defect classification model

As illustrated in Fig. 6, we first deploy the pre-trained SqueezeNet architecture to build a multi-stream CNN model, computing representative features on images captured under different lighting directions. Moreover, we update the parameters of SqueezeNet models to enhance the discriminative ability of feature extraction through deep supervision on defect classification tasks on individual streams. Finally, we integrate the channel attention (CA) mechanism to identify the most representative features under multiple illumination directions for the subsequent defect classification task.

3.1. SqueezeNet-based feature extraction and fusion

In recent years, numerous deep CNN models such as GoogLeNet [40], VGG [41] and ResNet [25] have been built to achieve high-accuracy image classification using large-scale labeled images (e.g., ImageNet [19]). However, it is impractical to capture a large number of sample images containing various types of defects to train deep CNN models. Therefore, our proposed classification model is built on the light-weight SqueezeNet architecture [26] which is easy to fine-tune and less prone to small dataset over-fitting.

As illustrated in Fig. 7, the SqueezeNet contains 9 consecutive fire modules for feature extraction. To reduce the number of parameters, the channel number of the squeeze convolutional layer in each fire module is purposely set to a quarter of the

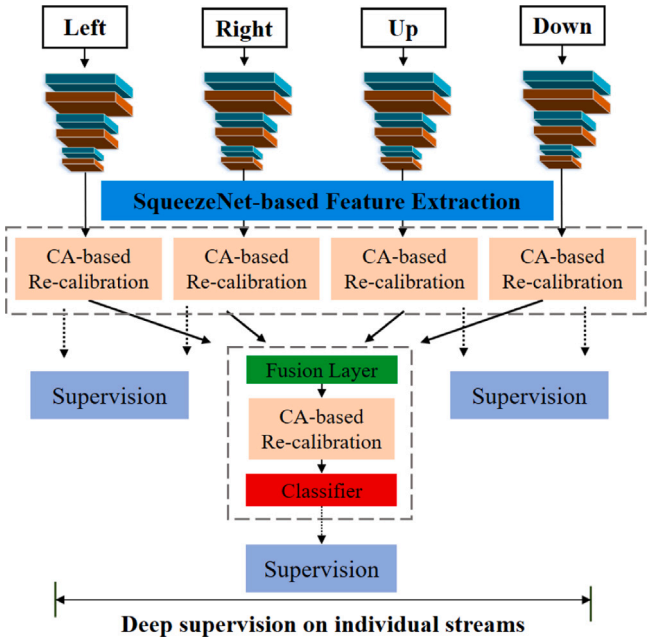

Fig. 6. The architecture of the proposed multi-stream CNN model to perform extraction and fusion of multi-light source illuminated features for high-accuracy surface defect classification.

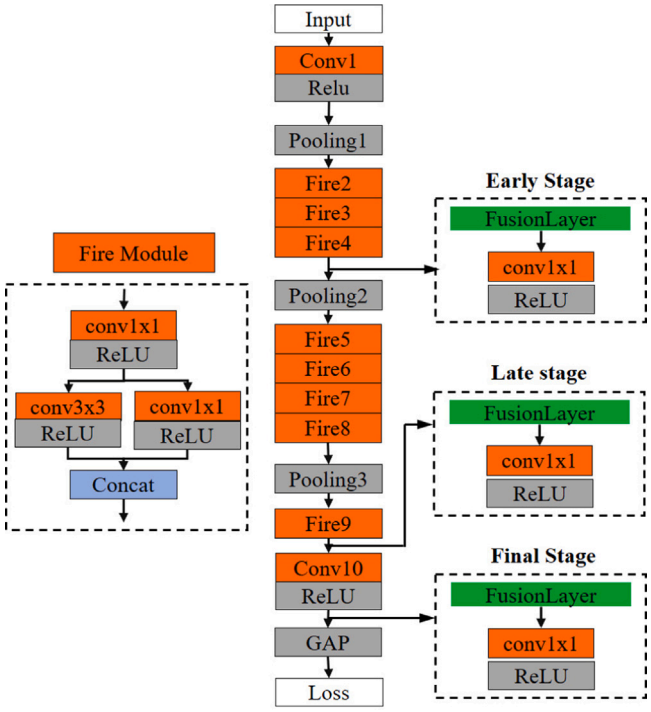

Fig. 7. The architecture of the pre-trained SqueezeNet model and features extracted in different convolutional stages.

expand layer. The channel number of Conv10 is set to 5 for classifying five categories of image samples (i.e., Normal, BrightLine, Deformation, Scratch, and Dent). We deploy a global average pooling (GAP) layer [21,24] to calculate the average values over the 13×13 slices to output $1 \times 1 \times 5$ matrix as the prediction result. The detailed network configurations of the SqueezeNet architecture (fire modules, activation layers, pooling layers, and a loss layer) are shown in Table 2.

Table 2
The detailed configurations of SqueezeNet model. The parameters of the filters are represented as $C \times W \times L$, where C, W, and L are the channel numbers, kernel width, and kernel length, respectively.

Layers	Output size	Filter size
Input	224 × 224 × 3	
Conv 1	109 × 109 × 96	96 × 7 × 7, stride 2
Pool 1	54 × 54 × 96	3 × 3 Pooling, stride 2
Fire 2	54 × 54 × 128	$16 \times 1 \times 1, 64 \times 1 \times 1, 64 \times 3 \times 3$
Fire 3	54 × 54 × 128	$16 \times 1 \times 1, 64 \times 1 \times 1, 64 \times 3 \times 3$
Fire 4	54 × 54 × 256	$32 \times 1 \times 1, 128 \times 1 \times 1, 128 \times 3 \times 3$
Pool 4	27 × 27 × 256	3 × 3 Pooling, stride 2
Fire 5	27 × 27 × 256	$32 \times 1 \times 1, 128 \times 1 \times 1, 128 \times 3 \times 3$
Fire 6	27 × 27 × 384	$48 \times 1 \times 1, 192 \times 1 \times 1, 192 \times 3 \times 3$
Fire 7	27 × 27 × 384	$48 \times 1 \times 1, 192 \times 1 \times 1, 192 \times 3 \times 3$
Fire 8	27 × 27 × 512	$64 \times 1 \times 1, 256 \times 1 \times 1, 256 \times 3 \times 3$
Pool 8	13 × 13 × 128	3 × 3 Pooling, stride 2
Fire 9	13 × 13 × 512	$64 \times 1 \times 1, 256 \times 1 \times 1, 256 \times 3 \times 3$
Conv 10	13 × 13 × 5	6 × 13 × 13, stride 1
Pool 10	1 × 1 × 5	13 × 13 Pooling

We deploy four pre-trained SqueezeNet models to compute representative features on images captured under four lighting directions (Left, Right, Up, and Down). A fusion layer is then applied to aggregate the computed four-stream CNN features as

$$\mathbf{M} = f(\mathbf{L}, \mathbf{R}, \mathbf{U}, \mathbf{D}), \tag{1}$$

where f is the fusion function, $\mathbf{L} \in \mathbb{R}^{C_1 \times H_1 \times W_1}$, $\mathbf{R} \in \mathbb{R}^{C_2 \times H_2 \times W_2}$, $\mathbf{U} \in \mathbb{R}^{C_3 \times H_3 \times W_3}$, and $\mathbf{D} \in \mathbb{R}^{C_4 \times H_4 \times W_4}$ are the features computed on individual images illuminated by the Left, Right, Up and Down light sources, respectively, and $\mathbf{M} \in \mathbb{R}^{C \times H \times W}$ is the fused multi-light source illuminated feature maps. C, H and W indicate the channel number, height and weight of feature maps, respectively. Since four feature extraction streams are built using the same SqueezeNet architecture, we have $C_1 = C_2 = C_3 = C_4$, $W_1 = W_2 = W_3 = W_4$, $H_1 = H_2 = H_3 = H_4$. We adopt the commonly-used Concatenation operation (f^{cat}) to integrate multi-light source illuminated features as

$$\begin{cases} \mathbf{M}^{cat}_{c,h,w} = \mathbf{L}_{c,h,w} \\ \mathbf{M}^{cat}_{C+c,h,w} = \mathbf{R}_{c,h,w} \\ \mathbf{M}^{cat}_{2\times C+c,h,w} = \mathbf{U}_{c,h,w} \\ \mathbf{M}^{cat}_{3\times C+c,h,w} = \mathbf{D}_{c,h,w} \end{cases}, \tag{2}$$

where $c \in (1, 2, \ldots, C)$ is the channel index of feature maps, $h \in (1, 2, \ldots, H)$ and $w \in (1, 2, \ldots, W)$ indicate the spatial locations. Note $\mathbf{M}^{cat} = f^{cat}(\mathbf{L}, \mathbf{R}, \mathbf{U}, \mathbf{D})$ stacks multi-light source illuminated features **L**, **R**, **U**, and **D** at the same position h, w but across the feature channels c. A 3×3 kernel size convolutional layer is added to the fusion layer to decrease the channel number of $\mathbf{M}^{cat}$ from $4C$ to C.

The SqueezeNet model deploys stacked fire modules and convolutional layers to generate representative feature maps in different convolutional stages. It is well known that high-level features extracted in deeper layers typically encode global semantic information while low-level features computed in shallower layers tend to characterize local textured details [42]. It remains an open question which type of features are more suitable for the defect classification task. As illustrated in Fig. 7, we experimentally evaluate the performances of CNN models using features extracted in different convolutional stages (early stage features of Fire4, late stage features of Fire9 and final stage features of Conv10) and discuss the optimal scheme for extracting and integrating multi-light source illuminated features to achieve accurate surface defect classification results in Section 4.2.1.

3.2. Deep supervision on individual streams

The backbone squeezeNet architecture is pre-trained using images in ImageNet for general object classification tasks (e.g., bird, flower, person, and vehicle) thus is not optimal for distinguishing surface defects on highly reflective metals. To perform better defect classification, we design a multi-term loss function to update the parameters of our proposed multi-stream CNN model. Given the features extracted on individual streams (**L**, **R**, **U**, **D**) and the fused multi-light source illuminated feature **M**, our proposed model generates multiple prediction results to incorporate deep supervision on defect classification task on individual streams. The multi-term loss $\mathcal{L}_F$ is formulated as

$$\mathcal{L}_F = (\alpha_{\mathbf{L}}\mathcal{L}_{\mathbf{L}} + \alpha_{\mathbf{R}}\mathcal{L}_{\mathbf{R}} + \alpha_{\mathbf{U}}\mathcal{L}_{\mathbf{U}} + \alpha_{\mathbf{D}}\mathcal{L}_{\mathbf{D}} + \alpha_{\mathbf{M}}\mathcal{L}_{\mathbf{M}}), \tag{3}$$

where the cross entropy loss terms $\mathcal{L}_{\mathbf{L}}$, $\mathcal{L}_{\mathbf{R}}$, $\mathcal{L}_{\mathbf{U}}$, and $\mathcal{L}_{\mathbf{D}}$ evaluate the intermediate prediction results of four individual streams (**L**, **R**, **U**, **D**), the cross entropy loss $\mathcal{L}_{\mathbf{M}}$ measures the correctness of the final prediction based on fused multi-light source illuminated feature. $\alpha_{\mathbf{L}}$, $\alpha_{\mathbf{R}}$, $\alpha_{\mathbf{U}}$, $\alpha_{\mathbf{D}}$ and $\alpha_{\mathbf{M}}$ are the weights of multiple loss terms which are empirically set to 1. Given the ground-truth label k and the prediction y, the cross entropy loss function $\mathcal{L}$ is defined as

$$\mathcal{L} = -\sum_{k=1}^{5} t_k[k \log \Pr(y = k) + (1 - k) \log(1 - \Pr(y = k))], \tag{4}$$

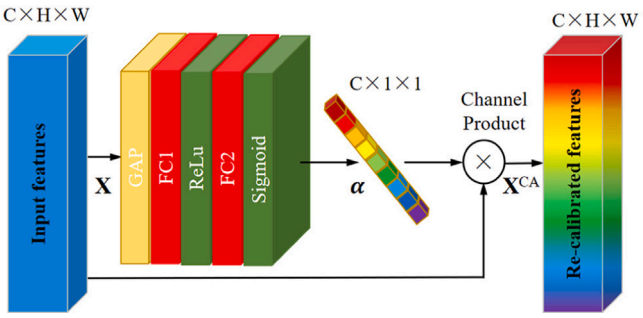

Fig. 8. The architecture of the proposed CA-based feature re-calibration.

where $t_k = 1$ when the label of input image is k, otherwise $t_k = 0$. $\Pr(y = k)$ measures the confidence of the prediction as

$$\Pr(y = k) = \frac{e^{G_k}}{\sum_{j=1}^{5} e^{G_j}}, \tag{5}$$

where G_k is the kth output of the GAP layer. Based on the proposed multi-term loss function $\mathcal{L}_F$, we update the parameters of the multi-stream CNN model to generate more discriminative features on individual streams for defect classification. Experiments are carried out to evaluate the effectiveness of incorporating deep supervision on individual streams in Section 4.2.2.

3.3. CA-based feature re-calibration

After computing multiple features on images captured under different lighting directions, we integrate the CA mechanism in the multi-stream CNN model to perform re-calibration of feature responses as illustrated in Fig. 8. Given an input feature $\mathbf{X} \in \mathbb{R}^{C\times H\times W}$, we firstly apply a GAP operation to squeeze the feature in spatial dimensions $H \times W$. A channel-wise feature tensor $z \in \mathbb{R}^{C\times 1\times 1}$ is calculated as

$$z(c) = GAP(\mathbf{X}) = \frac{1}{H \times W} \sum_{h=1}^{H} \sum_{w=1}^{W} \mathbf{X}(c, h, w), \tag{6}$$

where $\mathbf{X}(c, h, w)$ denotes the value at spatial location (h, w) of the cth channel of $\mathbf{X}$. A gating mechanism [43] is integrated to assign weights to the feature maps across different channels and it is composed of two fully-connected (FC) layers and a ReLU activation function as

$$\alpha = Sig(FC(Re(FC(z)))), \tag{7}$$

where $FC(\cdot)$ denotes the FC layers, $Re(\cdot)$ represents the ReLU function. A sigmoid function $Sig(\cdot)$ is introduced to project the weights to the range of 0~1, enhancing nonlinearity of network and ensuring the convergence of gradient transfer. The re-calculated output $\mathbf{X}^{CA}$ is obtained by rescaling the input feature $\mathbf{X}$ with the computed attention weights α as

$$\mathbf{X}^{CA}(c) = \alpha \cdot \mathbf{X}(c), \tag{8}$$

It is worth mentioning that the channel weights α are scene-specific and self-learned without any supervision or human intervention. Therefore, the CA mechanism can adaptively assign weights to different feature channels to enhance the informative features as well as to suppress redundant ones, performing the adaptive fusion of the most representative features. Note we apply the CA mechanism to re-calibrate features extracted in individual streams (**L**, **R**, **U**, **D**) as well as the fused one (**M**). We will experimentally evaluate the effectiveness of CA-based feature re-calibration in Section 4.2.2.

4. Experiments

4.1. Implementation details

All the experiments are implemented on the publicly accessible PyTorch framework. The SqueezeNet model [26] pre-trained in ImageNet is used to initialize the parameters of multi-stream feature extractors in the proposed model. The parameters of newly added or modified layers/modules, such as fusion layers and CA-based re-calibration modules, are initialized with Xavier distribution. The batch-size is set to 64 and the maximum training iteration is fixed to 4,000. The "Cosine Annealing" learning policy is used and the initial learning rate is set to 0.0025. The network is trained using Stochastic Gradient Descent (SGD) algorithm [44], momentum and weight decay are set to 0.9 and 0.0001, respectively. It takes less than 60 min to train our proposed CNN model on a NVIDIA Quadro P6000 GPU (24G RAM). In the inference phase, the prediction class which has the highest confidence score is considered as the classification result. These defect classifiers are assessed by computing the classification accuracy (%), meaning the percentage of correctly classified image numbers in each class [16,17,28,32].

Table 3
The classification accuracy(%) of models using images illuminated by single or multi-light sources.

	Normal	BrightLine	Deformation	Dent	Scratch	Average
Single-light source (Left)	83.4	78.6	89.3	85.0	83.7	80.0
Single-light source (Right)	88.6	83.8	89.2	94.7	83.1	87.5
Single-light source (Up)	72.1	85.2	67.8	73.7	79.7	74.3
Single-light source (Down)	86.3	80.5	69.8	93.9	75.1	80.7
Single-light source (Uniform)	80.0	81.7	69.9	83.0	87.3	79.8
Multi-light source	89.8	82.8	89.8	96.9	85.9	88.4

Table 4
The classification accuracy(%) of models using multi-light source illuminated features extracted in different stages.

	Normal	BrightLine	Deformation	Dent	Scratch	Average
Early stage (Fire4)	92.4	80.2	83.3	89.1	85.5	86.6
Late stage (Fire9)	92.1	85.0	90.8	96.5	89.6	90.4
Final stage (Conv10)	89.8	82.8	89.8	96.9	85.9	88.4

Table 5
The classification accuracy(%) of a number of network alternatives with/without performing individual stream deep supervision and CA-based feature re-calibration.

Model	Normal	BrightLine	Deformation	Dent	Scratch	Average
Baseline	92.1	85.0	90.8	96.5	89.6	90.4
Baseline+DS	94.2	92.9	89.7	96.1	92.6	93.0
Baseline+DS+CA	94.9	94.9	93.8	97.1	94.5	94.9

4.2. Performance analysis

We set up a number of experiments to evaluate the proposed multi-stream CNN model for fusion of multi-light source illuminated features and the effectiveness of two important techniques (individual stream deep supervision and CA-based feature re-calibration) to improve defect classification accuracy.

4.2.1. Multi-light source illuminated feature fusion

We show comparative results of different classification models based on images illuminated by single or multi-light sources in Table 3. Note the multi-stream CNN models are all built based on the pre-trained SqueezeNet architecture and further fine-tuned using images captured under single (feeding individual streams using single-light source illuminated images) or multiple lighting directions. The experimental results demonstrate that the fusion of complementary features extracted on multi-light source illuminated images provides an effective technique to generate more representative features of surface defects and thus can achieve more accurate classification results.

Our built SqueezeNet-based model deploys a number of stacked fire modules and convolutional layers in different convolutional stages to generate multi-scale feature maps. It is important to investigate whether the features computed in deeper or shallower layers are more suitable for defect classification. We experimentally evaluate the performances of multi-stream CNN models using features computed in the early stage (Fire4), late stage (Fire9), and final stage (Conv10). The comparative results in Table 4 clearly indicate that using features computed in deeper convolutional layers or fire modules generally leads to higher recognition accuracy (early stage 86.6% vs. late stage 90.4% vs. final stage 88.4%). Our experimental results suggest that semantic feature maps are more suitable for image-level analysis tasks such as target detection or classification. Another interesting observation is that performance using features of Conv10 is not satisfactory for edge-like defects such as BrightLine (late stage 85.0% vs. final stage 82.8%) and Scratch (late stage 89.6% vs. final stage 85.9%). A reasonable explanation is that features extracted in the final convolutional stage contain very limited low-level image cues and thus cannot well characterize local texture patterns and structural edges of BrightLine or Scratch defects.

4.2.2. Deep supervision and CA-based re-calibration

We propose to incorporate deep supervision on individual streams and CA-based feature re-calibration in our proposed multi-stream CNN model to improve defect classification accuracy. Ablation experiments are set up to evaluate their effectiveness. Based on the multi-stream CNN model using features extracted in late stage (Fire9), we built a number of network alternatives including (1) Baseline — using no deep supervision or CA-based re-calibration, (2) Baseline+DS — using deep supervision but no CA-based re-calibration, and (3) Baseline+DS+CA — using both deep supervision and CA-based re-calibration. The comparative results are shown in Table 5.

Based on the multi-term loss function $\mathcal{L}_F$ defined in Eq. (3), our proposed model not only measures the correctness of the final prediction based on the fused multi-stream features but also evaluates the intermediate prediction results of four individual streams. Such practice significantly increases the classification accuracy from 90.4% to 93.0%, proving the effectiveness of incorporating deep supervision to generate more discriminative features on individual streams and achieve more accurate defect classification.

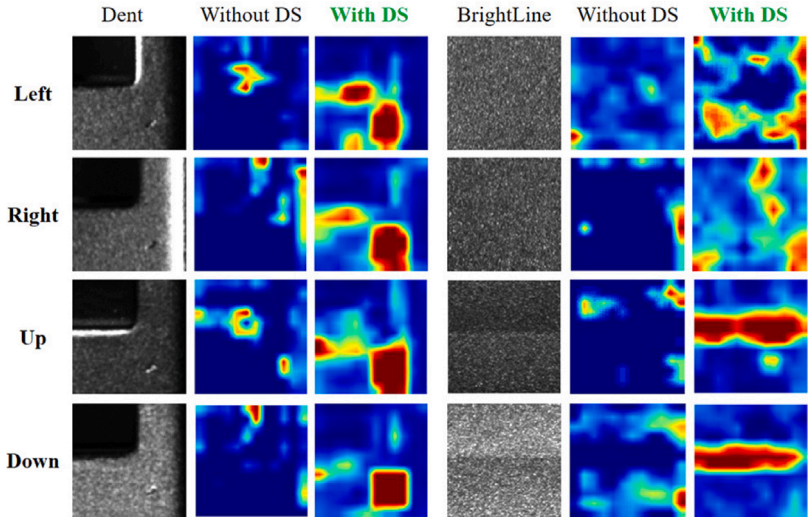

Fig. 9. The visualization of feature maps computed on sample images in the multi-light source illuminated dataset with/without deep supervision on individual streams.

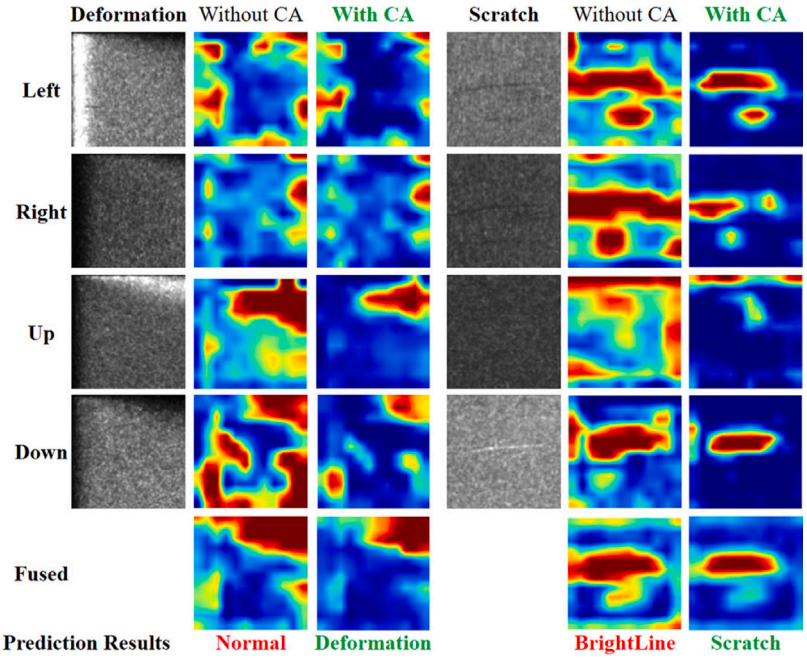

Fig. 10. The visualization of feature maps computed on sample images in the multi-light source illuminated dataset with/without CA-based re-calibration.

We further visualize the extracted features maps of individual streams with and without deep supervision in Fig. 9. The visualized feature map is obtained by multiplying the high-level feature maps with the channel-dependent GAP values and then adding up the weighted feature maps. It is observed that the computed feature maps without deep supervision are active on the non-defective regions with obvious structural edges. In comparison, incorporating deep supervision in the multi-stream CNN model can effectively enhance the ability of feature extraction on individual streams, successfully suppressing irrelevant features and generating more discriminative features on defective regions.

Given features extracted under individual lighting directions (**L**, **R**, **U**, **D**, **M**), we set larger weights for the informative features and smaller weights for redundant ones through the CA mechanism to perform scene-dependent re-calibration of feature responses. As the result, the classification accuracy is further increased from 93.0% to 94.9% as shown in Table 5. In Fig. 10, we illustrate the visualized features of two defect samples (Deformation and Scratch) with and without performing CA-based feature re-calibration. It is observed that the computed feature maps are heavily scattered and the defective regions cannot be correctly identified. Therefore,

Table 6
The classification accuracy(%) and model size of a number of deep-learning based surface defect classification models.

Model	Normal	BrightLine	Deformation	Dent	Scratch	Average	Model size
ETE [28]	79.9	91.6	88.7	98.0	87.9	86.3	2.5 MB
DECAF+MLR [17]	75.4	85.4	79.2	92.4	82.6	80.4	244.0 MB
SDC-SN-ELF+MRF [32]	87.1	87.9	83.2	94.5	87.0	87.6	3.1 MB
Proposed	94.9	94.9	93.8	97.1	94.5	94.9	3.5 MB

some defective samples (Deformation and Scratch) are misclassified as Normal and BrightLine, respectively. In comparison, the feature maps after CA-based re-calibration become visually more significant and concentrated on the defective regions and thus lead to correction classification of defect category. Quantitative and qualitative elevation results confirm the effectiveness of CA-based feature re-calibration to achieve more accurate defect detection/classification via tuning the feature responses towards the more representative channels under multiple lighting directions.

4.3. Comparisons with state-of-the-arts

Our proposed multi-stream model is compared with several best-performing CNN-based surface defect classification models. Note we only consider deep-learning-based classifiers including an end-to-end CNN model (ETE) [28], DECAF+MLR model [17], and SDC-SN-ELF+MRF model [32]. These models are either re-implemented using author-provided source codes or according to the original papers. For a fair comparison, these CNN models are trained using 11666 × 4 multi-light source illuminated images in the training dataset without applying any data augmentation techniques. In the testing phase, our proposed multi-stream CNN model takes four multi-light source illuminated images as input and directly generates a final prediction. In comparison, other single-stream CNN models [17,28,32] will generate four classification results based on four input images individually and produce the final prediction by assigning each image pixel to the class with the highest summed confidence score.

The classification accuracy and computational complexity of different CNN models are shown in Table 6. It is noted that the large-size DECAF+MLR model (244.0 MB), which utilizes the Decaf model [31] to extract features and trains the multi-linear regression (MLR) as the classifier, surprisingly performs the worst. The experimental results illustrate that it is critical to fine-tuning the parameters of pre-trained CNN models (e.g., Decaf or SqueezeNet) using training images of defects to achieve good performance on a new target domain. Also, the ETE model does not generate satisfactory results since its parameters are randomly initialized and cannot be adequately optimized using a small number of defect samples [17]. The SDC-SN-ELF+MRF model fine-tunes the pre-trained SqueezeNet model for the defect classification task and performs better than ETE and DECAF+MLR models. However, SDC-SN-ELF+MRF generates prediction results based on single-light source illuminated images and does not perform effective fusion of complementary features extracted on multiple streams. In comparison, our proposed multi-stream CNN model can effectively extract and integrate complementary features on images captured under different lighting directions. As the result, it is capable of generating significantly higher accuracy in comparison to some best-performing defect classification models [17,28,32]. On a single NVIDIA Quadro P6000 GPU (24G RAM), the proposed light-weight model (3.5 MB) can process more than 20 image patches of 200 × 200 pixels (covering the entire ROI of a workpiece) per second. Note it is much faster than a human inspector that typically will use 3–4 s to perform visual inspection of a target object including changing the lighting/viewing direction and making a prediction.

5. Conclusion

Most of the existing surface defect inspection methods deploy a single light source to illuminate the target workpieces thus cannot produce satisfactory results in practical industrial inspection tasks particularly for products made of highly reflective materials. In this paper, we present a multi-light source illumination/acquisition hardware system and a multi-stream CNN-based defect classification model for accurate and fast defect inspection on highly reflective surfaces based on the fusion of images illuminated by multi-light sources. Given the features extracted on individual streams, we incorporate deep supervision in our proposed multi-stream CNN model to enhance the discriminative ability of feature extraction in individual streams and utilize the channel attention (CA) mechanism to select the most discriminative features under multiple illumination directions. Experimental results validate the effectiveness of the proposed multi-stream CNN model, outperforming some best-performing defect classifiers and achieving significantly higher accuracy. Additionally, our proposed multi-stream CNN model could process over 20 input frames using a single NVIDIA Quadro P6000 GPU (24G RAM) and can be utilized for defect inspection of transparent or highly reflective objects which require multi-light source illumination.

In this paper, we only consider the illumination setup with four fixed lighting directions, which might not be optimal to stimulate and characterize insignificant surface defects. In the future, we plan to build a new surface defect illumination and acquisition system that is capable of capturing surface defect images under continuously changing lighting directions. Another noticeable drawback of the proposed method is that it only utilizes spatial features but ignores the important intensity variation between temporally adjacent frames. Therefore, we plan to design efficient 3D convolutional modules to process the captured image sequence and extract features in both spatial and temporal domains, thereby achieving better performance for defect detection, classification, and segmentation tasks.

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.

Acknowledgments

This research was supported by the National Natural Science Foundation of China (No. 52105526 and No. 52075485) and the National Key Research and Development Program of China (2020YFB1711400).

References

[1] Tian Wang, Yang Chen, Meina Qiao, Hichem Snoussi, A fast and robust convolutional neural network-based defect detection model in product quality control, Int. J. Adv. Manuf. Technol. 94 (9–12) (2018) 3465–3471.
[2] Hongbin Jia, Yi Lu Murphey, Jinajun Shi, Tzyy-Shuh Chang, An intelligent real-time vision system for surface defect detection, in: Pattern Recognition, 2004. ICPR 2004. Proceedings of the 17th International Conference on, vol. 3, IEEE, 2004, pp. 239–242.
[3] Yu Xie, Yutang Ye, Jing Zhang, Li Liu, Lin Liu, A physics-based defects model and inspection algorithm for automatic visual inspection, Opt. Lasers Eng. 52 (2014) 218–223.
[4] Santanu Ghorai, Anirban Mukherjee, M Gangadaran, Pranab K Dutta, Automatic defect detection on hot-rolled flat steel products, IEEE Trans. Instrum. Meas. 62 (3) (2013) 612–621.
[5] Chung-Feng Jeffrey Kuo, Chun-Yu Lai, Chih-Hsiang Kao, Chin-Hsun Chiu, Integrating image processing and classification technology into automated polarizing film defect inspection, Opt. Lasers Eng. 104 (2018) 204–219.
[6] P Kapsalas, P Maravelaki-Kalaitzaki, M Zervakis, ET Delegou, A Moropoulou, Optical inspection for quantification of decay on stone surfaces, NDT & E International 40 (1) (2007) 2–11.
[7] Chi-ho Chan, Grantham K.H. Pang, Fabric defect detection by Fourier analysis, IEEE Trans. Ind. Appl. 36 (5) (2000) 1267–1276.
[8] Yong-Ju Jeon, Doo-chul Choi, Sang Jun Lee, Jong Pil Yun, Sang Woo Kim, Defect detection for corner cracks in steel billets using a wavelet reconstruction method, J. Opt. Soc. Amer. A 31 (2) (2014) 227–237.
[9] Changhyun Park, Seho Choi, Sangchul Won, Vision-based inspection for periodic defects in steel wire rod production, Opt. Eng. 49 (1) (2010) 017202.
[10] Chaitali Tikhe, J.S. Chitode, Metal surface inspection for defect detection and classification using Gabor filter, Int. J. Innov. Res. Sci. Eng. Technol. 3 (6) (2014) 13702–13709.
[11] Roberto Medina, Fernando Gayubo, Luis M González-Rodrigo, David Olmedo, Jaime Gómez-García-Bermejo, Eduardo Zalama, José R Perán, Automated visual classification of frequent defects in flat steel coils, Int. J. Adv. Manuf. Technol. 57 (9–12) (2011) 1087–1097.
[12] Ehsan Amid, Sina Rezaei Aghdam, Hamidreza Amindavar, Enhanced performance for support vector machines as multi-class classifiers in steel surface defect detection, World Acad. Sci. Eng. Technol. 6 (7) (2012) 1096–1100.
[13] Maoxiang Chu, Rongfen Gong, Song Gao, Jie Zhao, Steel surface defects recognition based on multi-type statistical features and enhanced twin support vector machine, Chemometr. Intell. Lab. Syst. 171 (2017) 140–150.
[14] Daniel Soukup, Reinhold Huber-Mörk, Convolutional neural networks for steel surface defect detection from photometric stereo images, in: International Symposium on Visual Computing, Springer, 2014, pp. 668–677.
[15] J Mirapeix, PB García-Allende, A Cobo, OM Conde, JM López-Higuera, Real-time arc-welding defect detection and classification with principal component analysis and artificial neural networks, NDT & E Int. 40 (4) (2007) 315–323.
[16] Kechen Song, Yunhui Yan, A noise robust method based on completed local binary patterns for hot-rolled steel strip surface defects, Appl. Surf. Sci. 285 (21) (2013) 858–864.
[17] Ruoxu Ren, Terence Hung, Kay Chen Tan, A generic deep-learning-based approach for automated surface inspection, IEEE Trans. Cybern. 48 (3) (2018) 929–940.
[18] Alex Krizhevsky, Ilya Sutskever, Geoffrey E. Hinton, ImageNet Classification with deep convolutional neural networks, in: International Conference on Neural Information Processing Systems, 2012, pp. 1097–1105.
[19] Jia Deng, Wei Dong, Richard Socher, Li Jia Li, Kai Li, Fei Fei Li, ImageNet: A large-scale hierarchical image database, in: IEEE Conference on Computer Vision and Pattern Recognition, 2009, pp. 248–255.
[20] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, Alexander C. Berg, Li Fei-Fei, ImageNet Large scale visual recognition challenge, Int. J. Comput. Vis. 115 (3) (2015) 211–252, http://dx.doi.org/10.1007/s11263-015-0816-y.
[21] Alex Krizhevsky, Ilya Sutskever, Geoffrey E. Hinton, ImageNet Classification with deep convolutional neural networks, in: International Conference on Neural Information Processing Systems, 2012, pp. 1097–1105.
[22] Ross Girshick, Jeff Donahue, Trevor Darrell, Jitendra Malik, Rich feature hierarchies for accurate object detection and semantic segmentation, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2014, pp. 580–587.
[23] Sizhe Chen, Haipeng Wang, Feng Xu, Ya-Qiu Jin, Target classification using the deep convolutional networks for sar images, IEEE Trans. Geosci. Remote Sens. 54 (8) (2016) 4806–4817.
[24] Karen Simonyan, Andrew Zisserman, Very deep convolutional networks for large-scale image recognition, in: ICLR, 2015.
[25] Kaiming He, Xiangyu Zhang, Shaoqing Ren, Jian Sun, Deep residual learning for image recognition, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016, pp. 770–778.
[26] Forrest N. Iandola, Song Han, Matthew W. Moskewicz, Khalid Ashraf, William J. Dally, Kurt Keutzer, SqueezeNet: AlexNet-Level accuracy with 50x fewer parameters and <0.5MB] model size, 2016, ArXiv:1602.07360.
[27] Joseph Redmon, Santosh Divvala, Ross Girshick, Ali Farhadi, You only look once: Unified, real-time object detection, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016, pp. 779–788.
[28] Yi Li, Guangyao Li, Mingming Jiang, An end-to-end steel strip surface defects recognition system based on convolutional neural networks, Steel Res. Int. 88 (2) (2016).
[29] Junjie Xing, Minping Jia, A convolutional neural network-based method for workpiece surface defect detection, Measurement 176 (2021) 109185, http://dx.doi.org/10.1016/j.measurement.2021.109185.
[30] Dawei Li, Qian Xie, Xiaoxi Gong, Zhenghao Yu, Jinxuan Xu, Yangxing Sun, Jun Wang, Automatic defect detection of metro tunnel surfaces using a vision-based inspection system, Adv. Eng. Inform. 47 (2021) 101206, http://dx.doi.org/10.1016/j.aei.2020.101206.
[31] Jeff Donahue, Yangqing Jia, Oriol Vinyals, Judy Hoffman, Ning Zhang, Eric Tzeng, Trevor Darrell, Decaf: A deep convolutional activation feature for generic visual recognition, in: International Conference on Machine Learning, 2014, pp. 647–655.

[32] Guizhong Fu, Peize Sun, Wenbin Zhu, Jiangxin Yang, Yanlong Cao, Michael Ying Yang, Yanpeng Cao, A deep-learning-based approach for fast and robust steel surface defects classification, Opt. Lasers Eng. 121 (2019) 397–405.
[33] Jiangxin Yang, Guizhong Fu, Wenbin Zhu, Yanlong Cao, Yanpeng Cao, Michael Ying Yang, A deep learning-based surface defect inspection system using multi-scale and channel-compressed features, IEEE Trans. Instrum. Meas. (2020).
[34] Yanqi Bao, Kechen Song, Jie Liu, Yanyan Wang, Yunhui Yan, Han Yu, Xingjie Li, Triplet-graph reasoning network for few-shot metal generic surface defect segmentation, IEEE Trans. Instrum. Meas. 70 (2021) 1–11, http://dx.doi.org/10.1109/TIM.2021.3083561.
[35] Defu Zhang, Kechen Song, Jing Xu, Hongwen Dong, Yunhui Yan, An image-level weakly supervised segmentation method for no-service rail surface defect with size prior, Mech. Syst. Signal Process. 165 (2022) 108334.
[36] G. Rosati, G. Boschetti, A. Biondi, A. Rossi, Real-time defect detection on highly reflective curved surfaces, Opt. Lasers Eng. 47 (3–4) (2009) 379–384.
[37] Lin Li, Zhong Wang, Fangying Pei, Xiangjun Wang, Improved illumination for vision-based defect inspection of highly reflective metal surface, Chin. Opt. Lett. 11 (2) (2013) 021102.
[38] Awei Zhou, Bobo Ai, Pingge Qu, Wei Shao, On the defect detection for highly reflective rotary surface: an overview, Meas. Sci. Technol. (2020).
[39] A.A. Mohamed, R.V. Yampolskiy, Adaptive extended local ternary pattern (AELTP) for recognizing avatar faces, in: International Conference on Machine Learning and Applications, 2013.
[40] Christian Szegedy, Wei Liu, Yangqing Jia, Pierre Sermanet, Scott E. Reed, Dragomir Anguelov, Dumitru Erhan, Vincent Vanhoucke, Andrew Rabinovich, Going deeper with convolutions, in: 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2015, pp. 1–9.
[41] K. Simonyan, A. Zisserman, Very deep convolutional networks for large-scale image recognition, Comput. Sci. (2014).
[42] Hongyang Li, Jiang Chen, Huchuan Lu, Zhizhen Chi, Cnn for saliency detection with low-level feature integration, Neurocomput. 226 (2017) 212–220.
[43] Jie Hu, Li Shen, Gang Sun, Squeeze-and-excitation networks, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018, pp. 7132–7141.
[44] Léon Bottou, Stochastic gradient descent tricks, in: Neural Networks: Tricks of the Trade, Springer, 2012, pp. 421–436.