Pattern Recognition 109 (2021) 107571

Contents lists available at ScienceDirect

Pattern Recognition

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

CADN: A weakly supervised learning-based category-aware object detection network for surface defect detection

Jiabin Zhang[a,b], Hu Su[a,*], Wei Zou[a,b], Xinyi Gong[a], Zhengtao Zhang[a], Fei Shen[a]

[a] Institute of Automation, Chinese Academy of Sciences, Beijing 100190, China
[b] School of Artificial Intelligence, University of Chinese Academy of Sciences, Beijing 100049, China

ARTICLE INFO

Article history:
Received 26 January 2020
Revised 17 June 2020
Accepted 31 July 2020
Available online 1 August 2020

Keywords:
Weakly supervised learning
Automated surface inspection
Defect detection
Knowledge distillation

ABSTRACT

Large-scale data with human annotations is of crucial importance for training deep convolutional neural network (DCNN) to ensure stable and reliable performance. However, accurate annotations, such as bounding box and pixel-level annotations, demand expensive labeling efforts, which has prevented wide application of DCNN in industries. Focusing on the problem of surface defect detection, this paper proposes a weakly supervised learning method named Category-Aware object Detection network (CADN) to tackle the dilemma. CADN is trained with image tag annotations only and performs image classification and defect localization simultaneously. The weakly supervised learning is achieved by extracting category-aware spatial information in a classification pipeline. CADN could be equipped with either a lighter or a larger backbone network as the feature extractor resulting in better real-time performance or higher accuracy. To address the two conflicting objectives simultaneously, both of which are significant concerns in industrial applications, knowledge distillation strategy is adopted to force the learned features of a lighter CADN to mimic that of a larger CADN. Accordingly, the accuracy of the lighter CADN is improved while high real-time performance is maintained. The proposed approach is verified on our own defect dataset as well as on an open-source defect dataset. As demonstrated, satisfied performance is achieved by the proposed method, which could meet industrial requirements completely. Meanwhile, the method minimizes human efforts involved in image labelling, thus promoting the applications of DCNN in industries.

1. Introduction

Surface inspection is a significant concern in industrial production. Traditionally, surface inspection is performed manually. With the rapid development of image processing and deep learning technologies in recent years, it is expected that manual labor involved in the inspection could be replaced by automated manner not only to cut labor cost but also to improve the efficiency [1–3]. There is no doubt that DCNN has achieved great success in natural image tasks such as image classification [4], object detection [5] and semantic segmentation [6]. However, it is rarely used in industrial applications for the following reasons. Firstly, large-scale data with human annotations is required in the training process to ensure stable and reliable performance of DCNN. Unfortunately, accurate annotations, such as bounding box and pixel-level annotations, are time-consuming and commercially expensive to be obtained. This difficulty is particularly prominent in case of industrial applications due to the fact that the boundaries of weak defects on industrial images are vague and are hard to be accurately recognized even for experienced workers. Secondly, DCNN model involves complicated calculations and usually cannot meet the real-time requirement of industrial applications. Despite the high accuracy, the DCNN model cannot be applied.

Weakly supervised learning (WSL) provides an efficient means to reduce human efforts in image labelling by exploring alternative weak annotations. According to previous publications, weakly learning methods successfully performed object detection [7,8] or semantic segmentation task [9,10] with only image-level weak supervision. Besides, some attention-based models [11,12] select relevant regions to facilitate following decision procedure. Not intentionally, these models could achieve weakly learning tasks similarly as WSL methods. Among the above methods, several of them [7,9] adopt iterative training between two related networks leading to longer training time or even difficulty of convergence. For the others [8,10], additional time-consuming computation is involved in the extraction of spatial information, degrading the real-time performance. It should also be pointed out the low accuracy in object localization for the WSL methods. This is primarily driven by

[*] Corresponding author.
E-mail address: hu.su@ia.ac.cn (H. Su).

https://doi.org/10.1016/j.patcog.2020.107571
0031-3203/© 2020 Elsevier Ltd. All rights reserved.

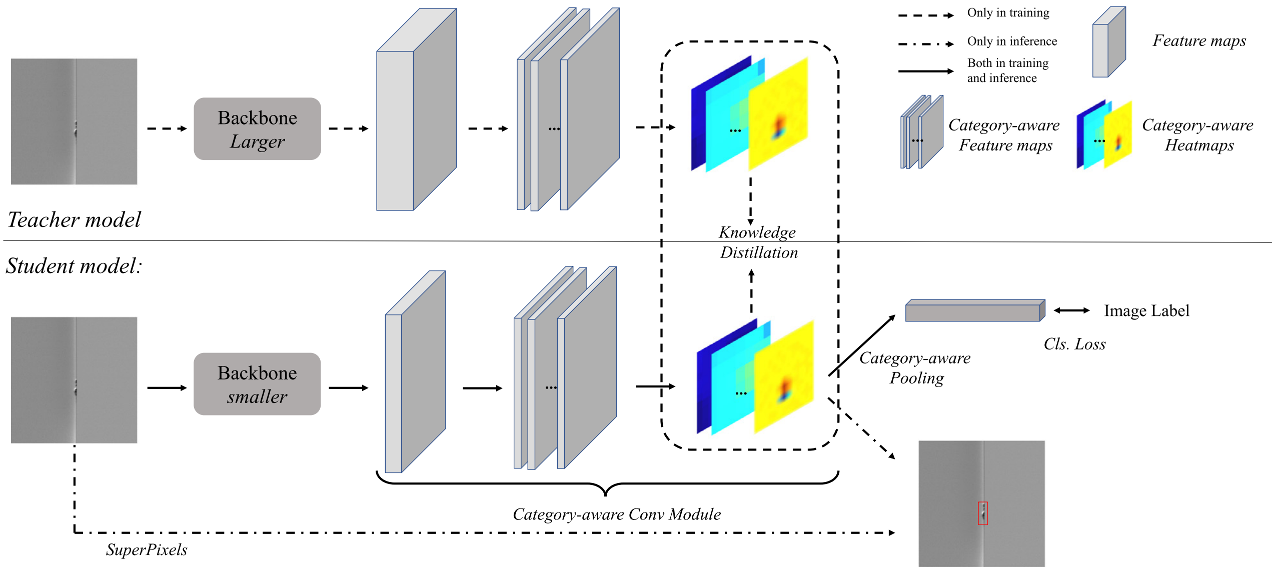

Fig. 1. The pipeline of the CADN Framework. The architecture of the proposed CADN network and the knowledge distillation strategy are illustrated.

the substantial reduction in resolution of feature map compared with original image. In WSL methods, feature map is the basis for spatial information extraction and thus its low resolution presents a negative impact on localization accuracy.

Actually, different strategies are proposed to increase the resolution of feature map, which, as pointed out above, is of great help for the WSL methods to improve localization accuracy. These strategies are summarized as follows. (1) The encoder-decoder structure [13], that captures context information in the encoder path and recovers high-resolution feature map in the decoder path. (2) Dilated convolution (a.k.a. atrous convolution) [6], based on which down-sampling operations, such as max pooling or strided convolution, are removed and thus the feature map resolution can be preserved. (3) Deconvolution (transposed convolution) [14], in which padding, interpolation and convolution are sequentially conducted on the feature map to increase the resolution at the end of a network. All the three methods are based on low-resolution feature map to generate high-resolution one. Although effective, spatial information missing is inevitable, which hinders further improvement in resolution. In addition to the above, the latest High-Resolution Network (HRNet) provides an efficient way to keep high-resolution feature maps for input image while passing throughout the network [15]. HRNet consists of multiple parallel branches with different resolutions. Lower resolution branches capture contextual information while higher resolution branches preserve spatial information. However, maintaining a high-resolution branch from beginning to end requires more computational cost.

Focusing on surface defect detection, a WSL method named **C**ategory-**A**ware object **D**etection network (CADN) is proposed in the paper. The method CADN performs image classification and defect localization simultaneously but only requires economical image tag annotations, which solves the difficulty of obtaining accurate annotations in industrial applications. In CADN, a novel Category-aware Conv-Pooling is proposed to explore weak image tag annotation, by which weakly supervised learning could be achieved. The module includes successive category-aware convolutions and category-aware pooling. The former category-aware convolutions extract spatial information for each category of object from the feature maps via a coarse-to-fine pipeline. The latter category-aware pooling transforms the category-aware spatial information into classification scores. By using the operations, iterative training and complex computations involved in previous WSL methods could be avoided. Meanwhile, to obtain high-resolution feature map, HRNet is adopted in CADN as the backbone. Naturally, CADN can achieve better accuracy or faster speed by equipping a large HRNet or a light HRNet, respectively. To address both the issues to make CADN more practical in industrial environment, knowledge distillation is utilized to force the outputs of a lighter CADN (student) to mimic the outputs of a larger CADN (teacher) in the training process of the student. Due to the additional supervision of the knowledge from the teacher model, the performance of the student model can be improved distinctly while its faster speed is maintained. The complete pipeline of CADN and the knowledge distillation strategy is illustrated in Fig. 1.

In conclusion, the main contributions of this paper can be described as follows:

1. A novel Category-aware Conv-Pooling module is proposed, which could explore weak image tag annotation to extract spatial information. And the latest HRNet is adopted to obtain high-resolution feature map to improve localization accuracy.

2. Knowledge distillation strategy is adopted to force the feature of a student CADN to mimic that of a teacher CADN, which contributes to the improvement in terms of both accuracy and speed.

3. As verified, weakly supervised defect detection is achieved and competitive results are obtained by using the proposed CADN method. The performance of CADN could fully meet industrial requirements.

The remainder of this paper is organized as follows. In Section II, we review the existing works related to our approach. In Section III, our proposed WSL object detection network CADN is introduced in details. Section IV describes the knowledge distillation strategy that improves the performance of the lighter and faster CADN. Section V provides the experimental results that verified our method. Finally, the paper is concluded in Section VI.

2. Related work

2.1. DCNN-based surface inspection

Compared with traditional methods, DCNN-based methods alleviate the difficulty of feature design while achieve improved accuracy, thus promoting progresses in image-related tasks greatly. In recent years, more and more attempts have been made to apply DCNN to surface inspection task to overcome the limitations of traditional methods [1,2]. A significant number of methods [16,17] accomplish the task by classifying normal and defect images. For example, MSPyrPool [16] is proposed to solve the steel defect classification problem on arbitrarily sized images. The network can be seen as a fully supervised hierarchical bag-of-features extension that is trained online and can be fine-tuned for any given task. However, it is worth mentioning that classification of the defect image cannot provide any size and location information for the defects. Wang et al. [18] design a joint detection CNN architecture that contains two major parts: the global frame classification part and the sub-frame detection part. The former learns to classify the whole image, and the later is implemented on the image patches generated by the sliding-window method. Ren et al. [19] propose a generic DCNN-based surface inspection approach. There are two phases in the proposed method. The first phase includes supervised training of patch classifier. The second phase uses the trained classifier on extracted patches, and generates the heatmap of whole image to predict the locations of defects. Based on the image partitioning operation, these two methods achieve defect localization roughly by using classification networks. Benefited from the great success of object detection algorithms applied in natural scenarios, Cha et al. [20] utilize Faster Region-based Convolutional Neural Network (Faster R-CNN) [5] to detect multiple types of damages accurately. Chen [21] et al. propose a cascade network to localize defects in a coarse-to-fine manner. The network includes two detectors to sequentially localize the cantilever joints and their fasteners and a classifier to diagnose the defects. However, their detectors need to be trained with sizeable datasets including 2366 and 6371 images, respectively. The bounding box annotations need to be labeled manually for each image. In this paper, a weakly supervised learning method is proposed to perform surface inspection, with which image classification and defect localization could be achieved simultaneously but only with the requirement of image tag annotations. The recent work, LEDNet [22], is similar to ours. However, LEDNet directly adopts the class activation mapping (CAM) technique [23] which is an exploring but less-powerful WSL method. In contrary, a novel Category-aware Conv-Pooling module is proposed in our paper, which is experimentally demonstrated to be more effective than the CAM. Another recent work, a weakly supervised network [24], is also proposed for surface defect detection. The proposed network is designed to be trained on a small number of images and achieves high classification accuracy. However, the method did not give a quantified detection results.

2.2. Weakly supervised learning

For a specific type of method, weakly supervised learning is treated as a multiple instance learning (MIL) problem by representing each image as a bag of instances (putative bounding boxes). The methods often involve iterative learning that alternates between training a regular-supervised model and selecting the most confident positive instances. Boxsup [7] completes the learning procedure by alternating between automatically generating region proposals and training a convolutional networks. Wang et al. [9] propose an iterative bottom-up and top-down framework that alternatively expands object regions and optimizes segmentation network. But frequently, these methods are susceptible to poor initialization or are with the difficulty of convergence. Several other methods extract or reserve higher-level features (e.g. spatial information generally, compared with class features) by designing additional time-consuming computations in the networks. CAM [23] is the first attempt for weakly supervised object detection. CAM utilizes the weights of the fully connective layer to obtain the weighted sum of feature maps to rebuilt the spatial features on feature maps. Oquab et al. [25] select the most informative region for the MIL prediction by Max pooling. ProNet [26] is a cascade of two networks, where the first generates bounding boxes and the second classifies them. Similarly, Bilen and Vedaldi [27] propose a specific architecture with two branches respectively dedicated to classification and detection. In [8], a weakly supervised region proposal network is proposed which is trained using only image-level annotations. Inspired by CAM and the method proposed in [25], a novel WSL method named CADN is proposed. CADN is an end-to-end network which is trained with image tag supervision to perform defect localization. The difference between CADN and the related methods is obvious and needs to be emphasized. CAM utilizes the weights of the fully connected layer to obtain the weighted sum of feature maps to rebuilt the spatial features. Then, the bounding boxes can be predicted on the rebuilt maps. In [25], the fixed size area (i.e. 224 × 224) corresponding to the maximum value among the n × m (i.e. 2 × 3) scores is treated as the bounding box of the object. In the training and the inference processes, multi scale images are needed and are fed to the network to detect objects with different scales. In CADN, based on the feature extracted by backbone network, Category-aware Conv and pooling are conducted. CADN directly predicts the bounding boxes on the category-aware heatmap. Compared with previous methods, the inference of CADN is straightforward and simple, meanwhile extra computations involved in CAM and the method in [25] could be avoided.

2.3. Knowledge distillation

Knowledge distillation referred as information transfer between different neural networks has been successfully exploited in many computer vision tasks. The research of exploring knowledge distillation in neural networks is pioneered by Hinton et al. in [28]. The authors investigate image classification problem to define the soft output of the teacher network as knowledge that contains useful information to represent intra-class similarity and inner-class diversity. Subsequently, FitNets [29] employs intermediate-level hints from hidden layers of the teacher network to train a thin and deep student network. In the study of object detection reported in [30], the small detection network is expected to learn more about object representation with the supervision of high-level features from the large networks. Wei et al. [31] utilize mimic and quantization strategies to train a very tiny detection network. In the method, mimic improves the performance of a student network by transferring knowledge from a teacher network. Meanwhile, quantization converts a full-precision network to a quantized one without large degradation of performance. By using distillation on unlabeled data, ADL [32] achieves better performance than the data distillation methods that simply utilize hard targets. By adopting ADL, the performance of the student detector excelled its teacher. Some researches [33,34] investigate the knowledge distillation strategy for training small semantic segmentation networks with the additional supervision from large networks. Liu et al. [34] simply view the segmentation problem as aggregated separate pixel classification problems and train compact semantic segmentation networks by the knowledge distillation strategy. He et al. [33] propose a new affinity distill module to transfer these long-rage dependencies among widely separated spatial regions from a teacher model to a student model. In our work, a knowledge distillation is developed to improve the detection accuracy of the lighter CADN while maintaining the high real-time performance. To the best of our knowledge, this is the first work to utilize knowledge distillation in surface inspection in industrial scenario.

3. The CADN framework

We propose a novel weakly supervised learning (WSL) framework named CADN for defect detection. As illustrated in Fig. 1,

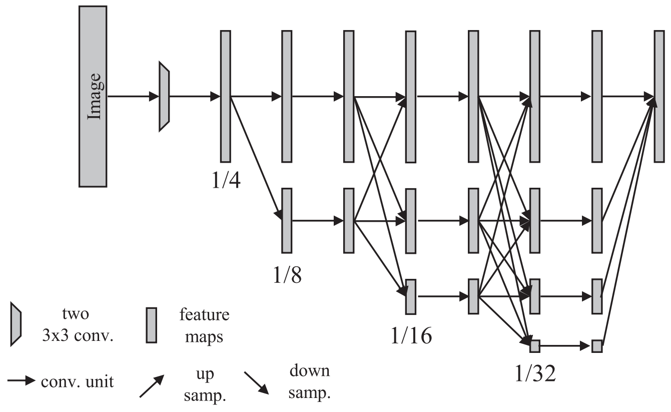

Fig. 2. Illustrating the architecture of the HRNet.

the pipeline of CADN network generally consists of a backbone network and a following Category-aware Conv-Pooling module. The backbone network extracts high-resolution features of the input image, and the following Category-aware Conv-Pooling module achieves weakly supervised object detection.

3.1. Backbone

High-resolution feature maps are obtained by using HRNet [15] as the backbone of our framework to improve the accuracy of defect localization. As Fig. 2 shows, HRNet starts with a high-resolution branch in the first stage. In every following stage, a new branch is added to current branches in parallel with resolution 1/2 of the lowest resolution in current branches. As the network has more stages, it will have more parallel branches with different resolutions and the resolutions from previous stages are all preserved in later stages.

In our framework, the backbone starts with two strided 3 × 3 convolutions which reduces the resolution of feature map to 1/4 of input image. Consequently, four successive stages are followed. The **1**st stage of HRNet contains 4 residual units where each unit is formed by a bottleneck structure with width (number of channels) of 64, followed by a 3 × 3 convolution reducing the width of feature maps to C. The **2**nd, **3**rd, **4**th stages contain 1, 4, and 3 multi-resolution blocks, respectively. The widths of the convolutions of the four brunches in the 4th stage are C, $2C$, $4C$, and $8C$, respectively. Each branch in the multi-resolution group convolution has 4 residual units and each unit has two 3 × 3 convolutions in each resolution. In the implementation, we set C to 18 and 32 for the lighter and the larger configuration which are called CADN-W18 and CADN-W32 respectively.

In the original paper [15] of HRNet, only the feature maps of the highest-resolution branch outputted by the 4th stage are used to predict the human keypoints. Unlike the task of human pose estimation which predicts the joint locations of the fixed size persons in [15], defect detection needs to predict all the defects with different scales. So we use all the outputs of four branches in the 4th stage by using upsampling to transform the resolution of other branches to the highest-resolution to obtain the concatenated feature maps $F \in \mathcal{R}^{\frac{W}{4}\times\frac{H}{4}\times 15C}$.

3.2. Category-aware Conv-Pooling module

The extraction and reserve of spatial information within feature maps plays a critical role in WSL object detection pipeline. CADN proposes Category-aware Conv-Pooling module to perform this task intuitively and graciously.

Table 1
The parameters of Category-aware Conv module.

	Number of group	Input channel	Output channel	Kernel size	Stride
$\boldsymbol{\phi_{c1}}$	1	$15C$	mK	3 × 3	1
ϕ_{c2k}	K	m	1	1 × 1	1

On the basis of the high-resolution feature maps $F \in \mathcal{R}^{\frac{W}{4}\times\frac{H}{4}\times 15C}$, as illustrated in Fig. 1, Category-aware Conv-Pooling module consists of two components: the Category-aware Conv module $\boldsymbol{\phi_c}$ and the Category-aware Pooling $\boldsymbol{P_c}$. The module of CADN is to transform the spatial feature maps to a classification vector so that the network can be trained under the supervision of image tag annotations. What's more important, the category-aware spatial information are extracted and reserved in the intermediate feature maps (also can be deemed as category-aware heatmaps).

3.2.1. Category-aware Conv module

Category-aware Conv module extracts and reserves spatial information via a coarse-to-fine pipeline. "Coarse" means each category of defect is assigned m-channels feature map to extract its spatial information:

$$F_{cat} = \boldsymbol{\phi_{c1}}(F \in \mathcal{R}^{\frac{W}{4}\times\frac{H}{4}\times 15C}) \tag{1}$$

where K is the number of defect category. $F_{cat} \in \mathcal{R}^{\frac{W}{4}\times\frac{H}{4}\times mK}$ can be seen as the category-aware features in which per extracted m channels features belonging to one-category defect.

Then "fine" means the spatial information of each category is transformed into a single-channel feature map which can be seen as category-aware heatmap:

$$H_{cat\,k} = \boldsymbol{\phi_{c2k}}(F_{cat\,k} \in \mathcal{R}^{\frac{W}{4}\times\frac{H}{4}\times m}) \tag{2}$$

where ϕ_{c2k} is the kth one of K groups convolutions $\boldsymbol{\phi_{c2}}$, which only belongs to the kth category defect. So that the m channels category-aware features of the kth category are transformed into a single-channel heatmap $H_{cat\,k} \in \mathcal{R}^{\frac{W}{4}\times\frac{H}{4}}$. The parameters of Category-aware Conv module are listed in Table 1 intuitively. Several samples of category-aware heatmaps are illustrated in Fig. 3. We can find that the spatial information of defect objects have been represented on the category-aware heatmaps.

With the obtained category-aware heatmaps $H_{cat} \in \mathcal{R}^{\frac{W}{4}\times\frac{H}{4}\times K}$, the region with maximum score for each category is extracted and used for point-wise localization. We used a common method in previous WSL methods [25,35], in which the category-aware heatmaps need to be threshold-ed to extract localization information of the object. Specifically, at first, the category-aware heatmaps are resized into the size of the input image. In the transformation of heatmaps to classification vector by the Category-aware Pooling, the classification score outputted by the pooling module is selected as the threshold for binarization of the resized heatmaps. Then, the bounding boxes can be obtained on the threshold-ed heatmaps. Moreover, to improve the accuracy of detection result, we adopt superpixels algorithm to modify the boundary of detected boxes as introduced in Section 3.3.

3.2.2. Category-aware Pooling

The Category-aware Pooling uses global average pooling operation to transform the category-aware heatmaps H_{cat} to a classification vector $Y \in \mathcal{R}^{K}$ as in Eq. (3). For Category-aware Pooling, we use global average pooling to transform the spatial activation of object into the classification score of the corresponding category. Global average pooling can extract more complete spatial feature than max-pooling, as well as avoid the influence of unexpected extreme activation.

$$Y = \boldsymbol{P_c}(H_{cat} \in \mathcal{R}^{\frac{W}{4}\times\frac{H}{4}\times K}) \tag{3}$$

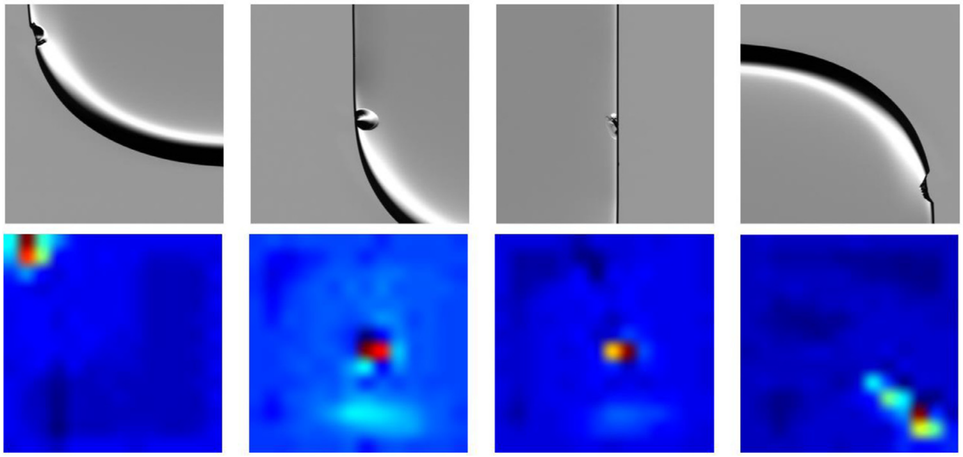

Fig. 3. Some sample results of category-aware heatmaps. The first row lists some images of a part of mobile phone cover-glass, which have defects on the edge. The second row lists corresponding category-aware heatmaps.

In this way, the CADN network can be trained under the supervision of global image labels $\hat{Y}_{cls}$ with Cross-Entropy loss:

$$\boldsymbol{L}_{cls} = -\sum_{k=1}^{K}[Y_k \log \hat{Y}_k + (1-Y_k) log(1-\hat{Y}_k)] \tag{4}$$

3.3. Superpixels post-processing

To improve the accuracy of detection result, we adopt superpixels algorithm to modify the boundary of detected boxes. The superpixels regions of each image are generated by the superpixels algorithm in [36]. For each side of a detection bounding box, a superpixels region which is located on this side and has the biggest intersection-over-union with the box is selected. Then this side of the box is extended to the boundary of the superpixels region.

4. Knowledge distillation for CADN

In industrial applications, both accuracy and speed are of crucial importance. Larger network could achieve high accuracy while smaller network could run faster at the cost of lower accuracy. In contrast to previous methods that attempt to make a better tradeoff between the two conflicting objectives, we adopt knowledge distillation strategy to overcome the dilemma. Knowledge distillation forces the outputs a faster CADN (with a smaller backbone as a student) to mimic the output of a high-performance CADN (with a larger backbone as a teacher) in the training process of the student model. By using the proposed strategy, the accuracy of the smaller network is improved while high real-time performance is maintained.

4.1. Teacher and student models

The student model has HRNet-18 as the backbone for feature extraction. To keep feature consistency between the teacher and the student models, HRNet-32 is selected as the backbone of the teacher model. As noted, HRNet-32 has the same architecture with HRNet-18 but is much larger.

4.2. Transfer module

With the determined teacher and student models, the key next step is knowledge representation and knowledge alignment between the two models. There are two choices in knowledge representation, feature map or heatmap. Heatmap is chosen as the defect knowledge based on the following two considerations. Firstly, the distillation on feature map would be computationally expensive due to the huge amount of channels. Secondly, the category-aware heatmap includes more explicit spatial information which can help the student model learn them more straightforwardly.

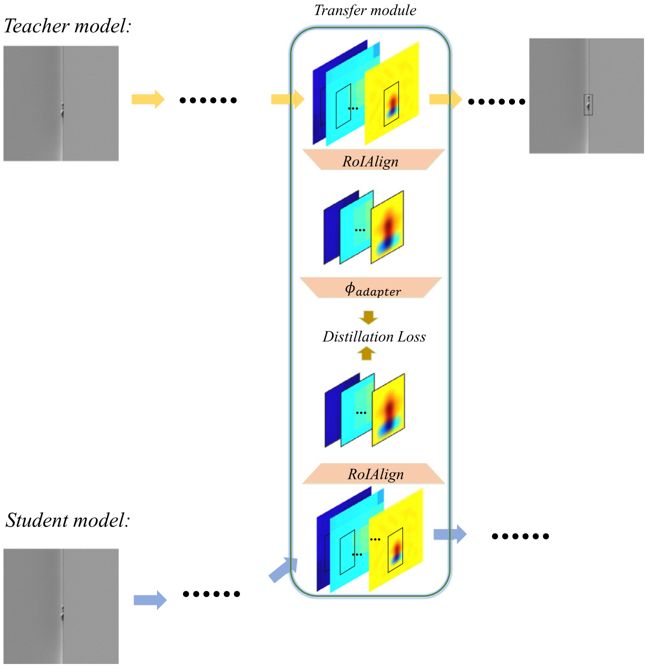

Fig. 4. The pipeline of knowledge distillation strategy.

As shown in Fig. 4, the transfer module performing knowledge distillation includes two components: the ROIAlign [37] and the convolutional adapter $\boldsymbol{\phi}_{\boldsymbol{adapter}}$. In distillation, we argue that only defect region needs to be considered to reduce calculation and improve distillation accuracy. The ROIAlign operation is adopted to extract the area of heatmaps of the teacher and the student models corresponding to defects detected by the techer model:

$$H_n^s = ROIAlign(H^s \in \mathcal{R}^{\frac{W}{4}\times\frac{H}{4}\times K}) \text{ by } bbox_n \tag{5}$$

where $H_n^s \in \mathcal{R}^{w\times h\times K}$, H^s is the heatmap of student model, $bbox_n$ is the bounding box of the nth defect. The same operation is also performed on the teacher model.

Then the convolutional adapter $\boldsymbol{\phi}_{adapter}$ is carried out on the extracted heatmap $H_n^{tk} \in \mathcal{R}^{w\times h\times K}$ of the teacher model to transfer the latent knowledge better. The mimicked teacher knowledge is calculated by:

$$H_n^{tk} = \boldsymbol{\phi}_{\boldsymbol{adapter}}(H_n^t \in \mathcal{R}^{w\times h\times K}) \tag{6}$$

4.3. Mimicking pipeline

Corresponding to Fig. 4, the whole knowledge strategy procedure is summarized in Algorithm 1. In the algorithm, a trained high-performance CADN-W32 network is needed in advance to act as the teacher model. In the entire training process, only the parameters of the student model are updated while those of the teacher model are fixed. In the **1**st stage, the student model CADN-W18 is trained under the image tag supervision. Consequently, the **2**nd stage is fine-tunning process where knowledge distillation is implemented and a different loss function is defined. The Mean-Squared Error (MSE) function is chosen as the distillation loss:

Algorithm 1 Knowledge distillation pipeline of CADN.

```
Require: The trained teacher model $T$ and the student model $S$.
Ensure: The trained student model $S$
  STAGE 1: Training student model CADN $S$:
      $W_S = \arg\min_{W_S} \boldsymbol{L}_{\mathrm{cls}}$
  STAGE 2: Fine-tune student model CADN $S$ for mimicking:
      $W_S = \arg\min_{W_S} (\alpha \boldsymbol{L}_{\mathrm{cls}} + \beta \boldsymbol{L}_{\mathrm{dis}})$
```

$$\boldsymbol{L}_{\mathrm{kd}} = \frac{1}{N}\frac{1}{K}\sum_{n=1}^{N}\sum_{k=1}^{K} \|H_n^s - H_n^{tk}\|_2^2 \tag{7}$$

where H_n^s and H_n^{tk} specify the student heatmaps and the transferred teacher heatmaps for the nth defect object, calculated by Eqs. (5) and (6), respectively.

By taking the advantage of the supervision of the teacher, the performance of the student can be improved straightforwardly while its faster speed has been maintained. The effectiveness of the knowledge distillation strategy has been verified in the ablation experiment in Section 5.3.3.

5. Experiments

5.1. Datasets and metrics

The proposed CADN network is experimentally verified on two datasets. The first dataset constructed by ourselves includes thousands of images of mobile phone cover glass (MPCG). The second one is the open-source defect dataset, DAGM. The two datasets are detailed as follows.

MPCG Dataset: The dataset is built for recognition of the edge crack defect of MPCG. The images in the dataset are collected from actual industrial productions and can be categorized into two classes: normal images and the ones with edge crack defect. The MPCG dataset contains 11,808 images with a resolution of 500 × 400 pixels, in which 5389 images are normal and 6419 images are with edge crack defect. To validate our methods better, we randomly split the whole dataset into training set and testing set with a ratio of 3:1. The specific numbers are shown in Table 2. The defect images have bounding box annotations for algorithm validation. Some images in the dataset are shown in the bottom row of Fig. 5.

Table 2
The image numbers of our own MPCG dataset.

	MPCG dataset	Training set	Testing set
Normal images	5389	4042	1341
Defect images	6419	4814	1605
Total images	11,808	8856	2952

DAGM Dataset: Images in this dataset are artificially generated but are similar to real world problems. DAGM consists of multiple data sets, each of which includes 1000 non-defect images and 150 images with labelled defects. The images in a single data set are similar, but different data sets are generated by using different texture models and defect models. Some images in the dataset are shown in the bottom row of Fig. 5.

Metric: As commonly suggested, we use the metrics, *Accuracy* and $F-measure$ to evaluate the performances of defective image classification, and use mean Average Precison (mAP) to evaluate the performances of defect localization. Classification *Accuracy* is calculated by dividing the number right-classified images of by the number of all testing images. Even a intuitive metric, *Accuracy* isn't sensitive and a true sense out of the evaluated method's performance when imbalanced class distribution occurs like in DAGM. So the $F-measur$ is adopted. The defective and normal images are respectively treated as positive and negative samples in classification task. Then the values of True positive (*TP*), False positive (*FP*), True negative (*TN*) and False negative (*FN*) can be counted, which represent the numbers of right-classified defective images, wrong-classified normal images, right-classified normal images and wrong-classified defective images. So *Precision* and *Recall* can be computed as:

$$Precison = \frac{TP}{TP + FP} \tag{8}$$

$$Recall = \frac{TP}{TP + FN} \tag{9}$$

The $F-measure$ for evaluating the classification performance can be defined as:

$$F - measure = \frac{(\gamma^2+1)*Precision*Recall}{\gamma^2*Precision+Recall} \tag{10}$$

In the comparison experiments, setting the parameter γ to 1, we get the $F1-measure$, which is the additional metric to obtain objective evaluations of classification performances of the methods.

The metric mAP is used to evaluate the defect detection results. For comparison fairness, a common threshold 0.5 is firstly selected and then, mAP is calculated by the manner adopted in Pascal VOC evaluation [38]. It means that, the detection bounding box is considered as TP if it has IoU $>$ 0.5 with one ground truth box.

5.2. Implementation details

CADN is implemented by PyTorch on a sever with four Nvidia GeForce GTX 1080 GPUs, running on Ubuntu 16.04 operating system. In the training process, SGD optimizer is utilized with the learning rate 0.02 and the batch size 16. We use a weight decay of 0.0001 and momentum of 0.9. The learning rate is dropped to 0.002 and 0.0002 at the 22nd and the 24th epoch, respectively. The iteration training is terminated after 26 epochs. The size of training image is set as 224 × 224. Parameter m in Category-aware Conv-Pooling module is set as 16. For knowledge distillation, α and β are both set as 0.5.

5.3. Ablation studies

To demonstrate the contributions of different parts of CADN, ablation experiments are conducted. In this section, ablation studies on high-resolution feature representation, Category-aware Conv-Pooling module, and knowledge distillation strategy are successively reported.

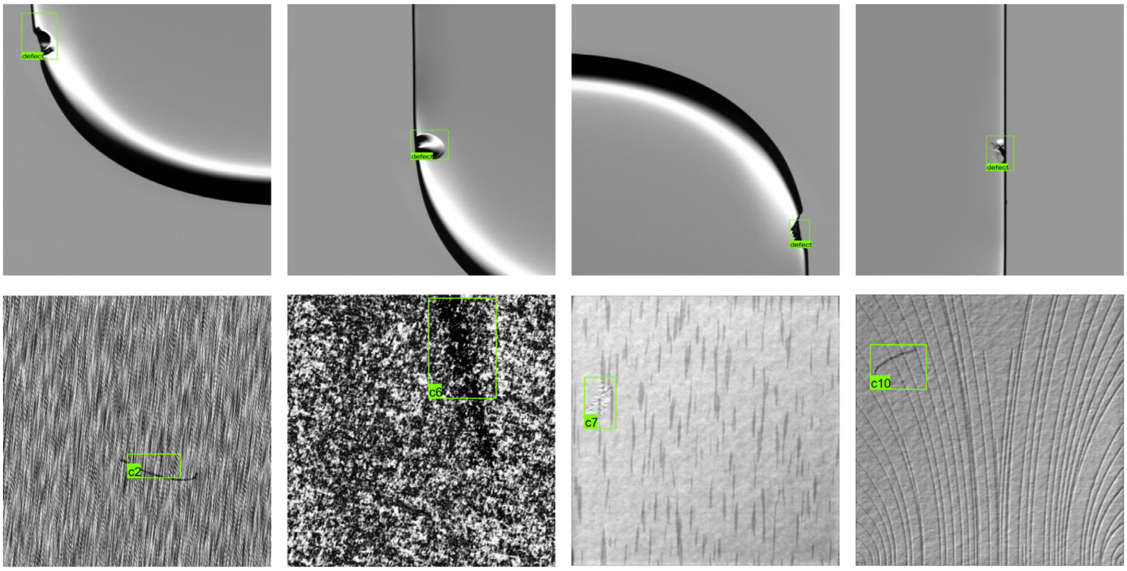

Fig. 5. Some sample results of CADN. The first row lists some defect localization results on MPCG Dataset. The second row lists defect localization results on DAGM Dataset.

Table 3
The performance of different backbones on the MPGC dataset. CADN achieves the highest performance by using HRNet-W32 than other backbones.

Backbones	Params	GFLOPs	*Cls.* Acc.	*Det.* mAP
VGG-11	133.96	7.83	82.9	46.9
ResNet-50	27.59	4.61	93.8	63.4
HRNet-W18	13.77	2.27	93.4	65.9
HRNet-W32	26.63	4.72	**98.7**	**68.5**

Table 4
The performance of CADN-W32 with different WSL components on the MPCG dataset. Our proposed Category-aware Conv-Pooling module achieves the highest metrics.

	Weakly Supervised Strategy	mAP
(a)	CAM [23]	55.1
(b)	Category-aware Conv module	60.2
(c)	Category-aware pooling	62.1
(d)	Category-aware Conv-Pooling	**68.5**

5.3.1. *Ablation study for high-resolution feature representation*

In this experiment, we evaluate the effectiveness of high-resolution feature representation by equipping different backbones for CADN: VGG-11, ResNet-50 and HRNet-W32. Among them, VGG-11 and ResNet-50 are commonly used backbones for visual recognition tasks and both of their feature maps have a 1/32 resolution of the input image. For the sake of fairness, knowledge distillation is not adopted, and parameters (Params) and FLOPs of the backbones are calculated with the input size of 224 × 224 to reflect the complexity of the backbones. As reported in Table 3, HRNet-W32 gets the highest *Accuracy* and mAP scores on the MPGC dataset among the three backbones, while its Params and GFLOPs are far lower than VGG-11 and comparable with ResNet-50. It can be concluded from Table 3 that HRNet-W32 achieves the best performance without any increase of memory space and computational cost. The superiority of the high-resolution feature representation is thus proved.

5.3.2. *Ablation study for Category-aware Conv-Pooling module*

The Category-aware Conv-Pooling module is the core of CADN. The module is designed to extract spatial information of foreground object within feature maps, based on which WSL is achieved. To verify the effectiveness, the following comparison experiments are conducted where four different WSL configurations are involved. (a) The WSL detection component in CAM [23]. It utilizes the weights of the fully connective layer to obtain the weighted sum of feature maps to generate the category-aware heatmap. (b) Only using Category-aware Conv module. Each category is assigned *m* channels feature map which is followed by a exclusive CAM component without Category-aware Pooling. (c) Only using Category-aware Pooling, which means a 1 × 1 convolutional layer replaces Category-aware Conv module to transfer feature maps to category-aware heatmap. (d) The proposed Category-aware Conv-Pooling module. As reported in Table 4, (d) achieves the best performance on the MPGC dataset among the different WSL configurations, proving the usefulness of the proposed Category-aware Conv-Pooling module.

5.3.3. *Ablation study for knowledge distillation*

Knowledge distillation and transfer module are of importance to improve the performance of CADN with a lighter backbone, which need to be emphasized and verified. Comparison experiments are conducted with the following different distillation strategies. (a) CADN-W18 without knowledge distillation. (b) Performing knowledge distillation on feature maps in CADN-W18. (c) Performing knowledge distillation on heatmaps without transfer module in CADN-W18. (d) Using *ROIAlign* without convolutional adapter $\phi_{adapter}$. (e) CADN-W18 with the proposed knowledge distillation strategy. The experimental results are listed in Table 5. As illustrated in the table, the strategy (e) can help (a) increase mAP by 1.3 (from 65.9 to 67.2) with the supervision of the teacher model, CADN-W32, which achieves the mAP of 68.5. Meanwhile,

Table 5
The performance of CADN-W18 with different knowledge distillation strategies on the MPCG dataset. Our proposed Category-aware Conv-Pooling module achieves the highest metrics.

	KD on *Feaure map*	KD on *Heatmap*	*ROIAlign*	$\phi_{adapter}$	mAP
(a)					65.9
(b)	✓				66.2
(c)		✓			66.5
(d)		✓	✓		66.8
(e)		✓	✓	✓	**67.2**

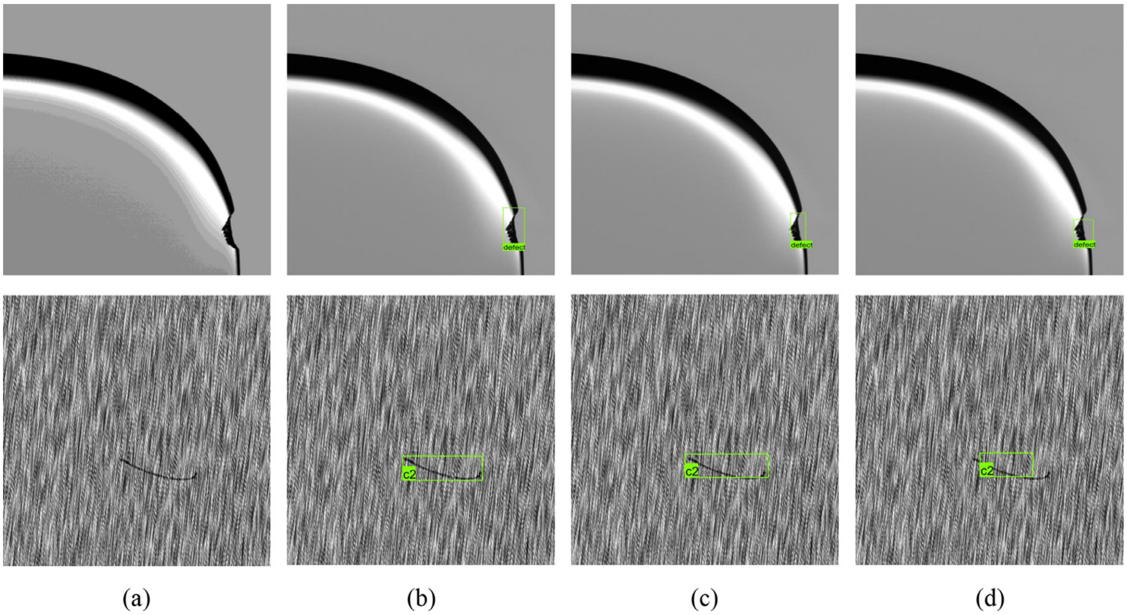

Fig. 6. The two images in (a) are from test sets of MPCG and DAGM Datasets. (b), (c) and (d) are their defect localization results by using Faster R-CNN, SSD and CADN-W32 respectively.

the increases of mAP by using other strategies (b) ~ (d) are all lower than 1.3. The superiority of the proposed knowledge distillation strategy is thus demonstrated. As listed in Table 3, the GFLOPs and the Params of HRNet-W18 are respectively 52% and 48% lower than HRNet-W32. Meanwhile, mAP achieved by HRNet-W18 is 2.6 lower than HRNet-W32. By using distillation strategy, the mAP gap is reduced to 1.3 (67.2 vs 68.5). That means that, nearly half of required memory space and computational cost could be saved at the expense of only a slight mAP degrade (1.3). This is of great significance for industrial applications. Therefore, knowledge distillation makes CADN more practical.

5.4. Comparison with the-state-of-the-art methods

For further demonstration, the proposed CADN method is compared with previous regular detection methods and WSL methods on the two dataset. Table 6 reports the results of classification and detection performances on the MPCG test set. CADN-W32 obtains classification accuracy of 98.1 and $F1-measure$ of 98.3 which greatly outperforms other weakly supervised (WS) methods and is comparable with the regular supervised (RS) methods. In the experiment, a universally acknowledged high-performance two-stage detector Faster R-CNN [5] with *RoIAlign* proposed in [37], a widely used single-stage detector SSD [39] and a recent single-stage detector FCOS [40] are selected as the comparison methods. These two RS detectors can obtain more accurate defect localization results than CADN as shown in Fig. 6 by being trained under the supervision of large quantities of manually labelled bounding box annotations. However, using only image-level annotations in training, CADN is more meaningful for practical applications even it achieves a slight lower mAP score than the RS methods as it dramatically reduces human burden in image labelling. Among the WSL methods, all of the CADNs achieve much higher mAP scores than CAM and CADN-W32 achieves the highest mAP score of 68.5, which is consistent with the analysis. Moreover, knowledge distillation strategy help CADN-W18 improve the *Accuracy* and mAP by 1.7 and 1.3 respectively without any increase of memory space and computational cost. We argue that such an improvement is significant in applying CADN-W18 in the industrial environment.

Notably, these regular detectors are originally designed for natural image detections while the paper focuses on industrial image. We can perceive that the two types of images are of discrepancy. The boundaries of foreground objects on natural images are clear while the boundaries of defects on industrial images are vague and are difficult to be accurately recognized. In response to that, we evaluate the impact of a lower IoU threshold in mAP calculations. Table 7 reports the results with the IoU threshold 0.3. As can be seen, the gap between CADN and regular supervised detectors has narrowed to about 5 from about 10. We think that, CADN has the capacity of defect recognition similar with regular detectors. However, the regular detectors could learn to accurately locate defect's boundaries with bounding box supervision while CADN could not. The threshold 0.5 is a common choice in regular detectors and is selected in the comparison experiments in Tables 6 and 8 for fairness.

Table 8 reports the results of the methods on DAGM test set. Despite the highest classification accuracy of 99.95, detection task cannot be accomplished by the method [41] and explicit information about defect, i.e. location and size, cannot be provided.

Table 6
Classification and detection performance on the MPCG test set.

Method	Type	*Cls.* Acc.	*Cls.* Pre.	*Cls.* Rec.	$F1-M.$	*Det.* mAP
SSD	RS	97.8	97.2	98.6	98.0	71.8
FCOS	RS	98.2	97.7	99.1	98.4	74.3
Faster R-CNN-50	RS	98.9	98.3	99.7	99.0	78.8
CAM	WS	88.9	86.7	94.0	90.2	55.1
CADN-W18	WS	93.4	92.4	95.8	94.1	65.9
CADN-W18(KD)	WS	95.1	94.4	96.8	95.6	67.2
CADN-W32	WS	98.1	97.7	98.8	98.3	68.5

Table 7
Detection results with different the IoU thresholds on the MPCG dataset.

Method	Type	regular *Det.* mAP	mAP with 0.3 IoU threshold
FCOS	RS	74.3	78.3
Faster R-CNN-50	RS	78.8	80.1
CADN-W32	WS	68.5	75.1

Table 8
Classification and Detection performance on DAGM defect test dataset.

Method	Type	*Cls.* Acc.	*Cls.* Pre.	*Cls.* Rec.	$F1 - M.$	*Det.* mAP
SSD	RS	88.2	53.0	90.1	66.8	65.2
FCOS	RS	88.6	54.0	91.4	67.9	68.9
Faster R-CNN-50	RS	89.8	57.0	91.6	70.2	68.9
Kim et al. [41]	RS	99.95	-	-	-	-
Staar et al. [42]	WS	83.0	-	-	-	-
CADN-W18	WS	86.2	48.7	90.0	63.2	56.8
CADN-W18(KD)	WS	87.6	51.6	90.7	65.8	58.3
CADN-W32	WS	89.1	55.1	92.0	69.0	61.2

Compared with other regular supervised detection methods, CADN-W32 achieves comparable $F1 - measure$ of 69.0 and has a gap of 7.7 with Faster R-CNN on mAP. The result is consistent with that of experiments carried out on the MPCG dataset. It should be also noted that, on such a imbalanced test set which includes 454 defective images and 2996 normal images, the metric *Recall* is more important for industries applications. As shown in Table 8, *Recall* of CADN-W32 is even higher than Faster R-CNN which further validates our method. In summary, the experiments on the two datasets demonstrate that, the proposed CADN is effective and can adapt to different industrial applications.

6. Conclusions

In this paper, we focus on the surface inspection task for which a weakly supervised learning method named CADN is proposed. Image classification and defect localization could be simultaneously achieved by CADN trained under image tag supervisions. A knowledge distillation strategy is adopted to improve the accuracy of the lighter CADN while maintaining its high real-time performance. Therefore, human efforts in image labelling, accuracy and speed are simultaneously considered in CADN, making the method practical in industrial applications. An MPCG dataset is constructed by collecting images from actual industrial productions and is employed together with the open source defect dataset DAGM to verify the proposed CADN method. Comparison and ablation experiments sufficiently demonstrate the effectiveness and superiority of CADN. Based on our work, further research should be continuous improvement of the lighter CADN. At present, the performance of the lighter CADN is improved by the knowledge distillation strategy, but still lags behind that of the larger CADN. Further improvement on the detection accuracy of the lighter CADN to approach or even exceed that of the larger CADN would be meaningful. Additionally, extending CADN to more general application scenarios would also be a promising research topic.

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.

Acknowledgements

The authors would like to thank the editors and the anonymous reviewers for their valuable comments and suggestions, which led to a substantial improvement of this paper. This work is supported by the National Key Research and Development Program of China under Grant 2018YFB1306303, and in part by the National Natural Science Foundation of China under Grant 61773374 and 61702323, and in part by the Major Basic Research Projects of Natural Science Foundation of Shandong Province under Grant ZR2019ZD07.

References

[1] C.S. Tsang, H.Y. Ngan, G.K. Pang, Fabric inspection based on the Elo rating method, Pattern Recognit. 51 (2016) 378–394.

[2] S.H. Hanzaei, A. Afshar, F. Barazandeh, Automatic detection and classification of the ceramic tiles surface defects, Pattern Recognit. 66 (2017) 174–189.

[3] Y. Yan, S. Kaneko, H. Asano, Accumulated and aggregated shifting of intensity for defect detection on micro 3d textured surfaces, Pattern Recognit. 98 (2020) 107057.

[4] K. He, X. Zhang, S. Ren, J. Sun, Deep residual learning for image recognition, in: IEEE Conference on Computer Vision and Pattern Recognition, 2016, pp. 770–778.

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

[6] L.-C. Chen, G. Papandreou, I. Kokkinos, K. Murphy, A.L. Yuille, DeepLab: semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected CRFs, IEEE Trans. Pattern Anal. Mach. Intell. 40 (4) (2018) 834–848.

[7] J. Dai, K. He, J. Sun, BoxSup: exploiting bounding boxes to supervise convolutional networks for semantic segmentation, in: IEEE International Conference on Computer Vision, 2015, pp. 1635–1643.

[8] P. Tang, X. Wang, A. Wang, Y. Yan, W. Liu, J. Huang, A. Yuille, Weakly supervised region proposal network and object detection, in: European Conference on Computer Vision, 2018, pp. 352–368.

[9] X. Wang, S. You, X. Li, H. Ma, Weakly-supervised semantic segmentation by iteratively mining common object features, in: IEEE Conference on Computer Vision and Pattern Recognition, 2018, pp. 1354–1362.

[10] S. Kwak, S. Hong, B. Han, Weakly supervised semantic segmentation using superpixel pooling network, in: AAAI Conference on Artificial Intelligence, 2017.

[11] K. Xu, J. Ba, R. Kiros, K. Cho, A. Courville, R. Salakhudinov, R. Zemel, Y. Bengio, Show, attend and tell: neural image caption generation with visual attention, in: International Conference on Machine Learning, 2015, pp. 2048–2057.

[12] J. Zhang, S.A. Bargal, Z. Lin, J. Brandt, X. Shen, S. Sclaroff, Top-down neural attention by excitation backprop, Int. J. Comput. Vis. 126 (10) (2018) 1084–1102.

[13] V. Badrinarayanan, A. Kendall, R. Cipolla, SegNet: a deep convolutional encoder-decoder architecture for image segmentation, IEEE Trans. Pattern Anal. Mach. Intell. 39 (12) (2017) 2481–2495.

[14] M.D. Zeiler, D. Krishnan, G.W. Taylor, R. Fergus, Deconvolutional networks, in: IEEE Conference on Computer Vision and Pattern Recognition, IEEE, 2010, pp. 2528–2535.

[15] K. Sun, B. Xiao, D. Liu, J. Wang, Deep high-resolution representation learning for human pose estimation, in: IEEE Conference on Computer Vision and Pattern Recognition, 2019, pp. 5693–5703.

[16] J. Masci, U. Meier, G. Fricout, J. Schmidhuber, Multi-scale pyramidal pooling network for generic steel defect classification, in: International Joint Conference on Neural Networks, 2013, pp. 1–8.

[17] V. Natarajan, T.-Y. Hung, S. Vaikundam, L.-T. Chia, Convolutional networks for voting-based anomaly classification in metal surface inspection, in: IEEE International Conference on Industrial Technology, 2017, pp. 986–991.

[18] T. Wang, Y. Chen, M. Qiao, H. 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.

[19] R. Ren, T. Hung, K.C. Tan, A generic deep-learning-based approach for automated surface inspection, IEEE Trans. Cybern. 48 (3) (2017) 929–940.

[20] Y.-J. Cha, W. Choi, G. Suh, S. Mahmoudkhani, O. Büyüköztürk, Autonomous structural visual inspection using region-based deep learning for detecting multiple damage types, Comput.-Aided Civ. Infrastruct. Eng. 33 (9) (2018) 731–747.

[21] J. Chen, Z. Liu, H. Wang, A. Núñez, Z. Han, Automatic defect detection of fasteners on the catenary support device using deep convolutional neural network, IEEE Trans. Instrum. Meas. 67 (2) (2017) 257–269.

[22] H. Lin, B. Li, X. Wang, Y. Shu, S. Niu, Automated defect inspection of led chip using deep convolutional neural network, J. Intell. Manuf. 30 (6) (2019) 2525–2534.

[23] B. Zhou, A. Khosla, A. Lapedriza, A. Oliva, A. Torralba, Learning deep features for discriminative localization, in: IEEE Conference on Computer Vision and Pattern Recognition, 2016, pp. 2921–2929.

[24] X. Liang, L. Shuai, D. Yong, L. Xiuxi, A weakly supervised surface defect detection based on convolutional neural network, IEEE Access (2020).

[25] M. Oquab, L. Bottou, I. Laptev, J. Sivic, Is object localization for free?-weakly-supervised learning with convolutional neural networks, in: IEEE Conference on Computer Vision and Pattern Recognition, 2015, pp. 685–694.

[26] C. Sun, M. Paluri, R. Collobert, R. Nevatia, L. Bourdev, ProNet: learning to propose object-specific boxes for cascaded neural networks, in: IEEE Conference on Computer Vision and Pattern Recognition, 2016, pp. 3485–3493.

[27] H. Bilen, A. Vedaldi, Weakly supervised deep detection networks, in: IEEE Conference on Computer Vision and Pattern Recognition, 2016, pp. 2846–2854.

[28] G. Hinton, O. Vinyals, J. Dean, Distilling the knowledge in a neural network, Stat 1050 (2015) 9.

[29] A. Romero, N. Ballas, S.E. Kahou, A. Chassang, C. Gatta, Y. Bengio, FitNets: hints for thin deep nets, arXiv:1412.6550 (2014).

[30] Q. Li, S. Jin, J. Yan, Mimicking very efficient network for object detection, in: IEEE Conference on Computer Vision and Pattern Recognition, 2017, pp. 6356–6364.

[31] Y. Wei, X. Pan, H. Qin, W. Ouyang, J. Yan, Quantization mimic: towards very tiny CNN for object detection, in: European Conference on Computer Vision, 2018, pp. 267–283.

[32] S. Tang, L. Feng, W. Shao, Z. Kuang, W. Zhang, Y. Chen, Learning efficient detector with semi-supervised adaptive distillation, arXiv:1901.00366 (2019).

[33] T. He, C. Shen, Z. Tian, D. Gong, C. Sun, Y. Yan, Knowledge adaptation for efficient semantic segmentation, in: IEEE Conference on Computer Vision and Pattern Recognition, 2019, pp. 578–587.

[34] Y. Liu, K. Chen, C. Liu, Z. Qin, Z. Luo, J. Wang, Structured knowledge distillation for semantic segmentation, in: IEEE Conference on Computer Vision and Pattern Recognition, 2019, pp. 2604–2613.

[35] A.J. Bency, H. Kwon, H. Lee, S. Karthikeyan, B. Manjunath, Weakly supervised localization using deep feature maps, in: European Conference on Computer Vision, 2016, pp. 714–731.

[36] M. Van den Bergh, X. Boix, G. Roig, B. de Capitani, L. Van Gool, SEEDS: Superpixels extracted via energy-driven sampling, in: European Conference on Computer Vision, 2012, pp. 13–26.

[37] K. He, G. Gkioxari, P. Dollár, R. Girshick, Mask R-CNN, in: IEEE International Conference on Computer Vision, 2017.

[38] M. Everingham, L. Van Gool, C.K. Williams, J. Winn, A. Zisserman, The pascal visual object classes (VOC) challenge, Int. J. Comput. Vis. 88 (2) (2010) 303–338.

[39] W. Liu, D. Anguelov, D. Erhan, C. Szegedy, S. Reed, C.-Y. Fu, A.C. Berg, SSD: Single shot multibox detector, in: European Conference on Computer Vision, 2016, pp. 21–37.

[40] Z. Tian, C. Shen, H. Chen, T. He, FCOS: Fully convolutional one-stage object detection, in: Proceedings of the IEEE International Conference on Computer Vision, 2019, pp. 9627–9636.

[41] S. Kim, W. Kim, Y.-K. Noh, F.C. Park, Transfer learning for automated optical inspection, in: International Joint Conference on Neural Networks, 2017, pp. 2517–2524.

[42] B. Staar, M. Lütjen, M. Freitag, Anomaly detection with convolutional neural networks for industrial surface inspection, Procedia CIRP 79 (2019) 484–489.

Jiabin Zhang received the B.Sc. degree in measurement and control technology and instrument from Shandong University, Jinan, China, in 2016. He is currently pursuing the Ph.D. degree in control theory and control engineering with the Institute of Automation, Chinese Academy of Sciences, Beijing, China, and also with the School of Artifficial Intelligence, University of Chinese Academy of Sciences, Beijing. His current research interests include industrial appearance inspection, computer vision, and weakly supervised learning systems for computer vision tasks.

Hu Su received the B.Sc. and M.Sc. degrees in information and computation science from Shandong University, Jinan, China, in 2007 and 2010, respectively, and the Ph.D. degree in control science and engineering from the State Key Laboratory of Management and Control for Complex Systems, Institute of Automation, Chinese Academy of Sciences (IACAS) Beijing, China, in 2013. Since 2013, he has been with IACAS, where he is currently an Associate Researcher with the Research Center of Precision Sensing and Control. His current research interests include intelligent control and optimization, and computer vision.

Wei Zou received the B.Sc. degree in control theory and control engineering from the Inner Mongolia University of Science and Technology, Baotou, China, in 1997, the M.Sc. degree in control theory and control engineering from Shandong University, Jinan, China, in 2000, and the Ph.D. degree in control theory and control engineering from the Institute of Automation, Chinese Academy of Sciences, Beijing, China, in 2003. Since 2012, he has been a Professor with the Research Center of Precision Sensing and Control, Institute of Automation, Chinese Academy of Sciences. His research interests mainly focus on visual control and intelligent robots.

Xinyi Gong received the B.Sc. degree in automation from Tsinghua University, Beijing, China, in 2014, and the Ph.D. degree in control science and engineering from the Institute of Automation, Chinese Academy of Sciences, Beijing, China, and the University of the Chinese Academy of Sciences, Beijing, in 2019. He is currently an research assistant with the Research Center of Precision Sensing and Control, IACAS. His current research interests include computer vision, image processing, pattern recognition, and machine learning.

Zhengtao Zhang received the B.Sc. degree in control science and engineering from the China University of Petroleum, Dongying, China, in 2004, the M.Sc. degree in control science and engineering from the Beijing Institute of Technology, Beijing, China, in 2007, and the Ph.D. degree in control science and engineering from the Institute of Automation, Chinese Academy of Sciences (IACAS), Beijing, in 2010. He is a Professor with the Research Center of Precision Sensing and Control, IACAS. His research interests include visual measurement, micro-assembly, and automation.

Fei Shen received the B.Sc. and M.Sc. degrees from Xidian University, China, and Beijing Institute of Technology, China, in 2007 and 2009, respectively, and the Ph.D. degree in control science and engineering from IACAS, China in 2012. He is currently an associate professor in the Research Center of Precision Sensing and Control, IACAS, China. His research interests include robot control, robot vision control and micro-assembly.