# An Efficient Convolutional Neural Network Model Based on Object-Level Attention Mechanism for Casting Defect Detection on Radiography Images

Chuanfei Hu and Yongxiong Wang

***Abstract*—Automatic detection of casting defects on radiography images is an important technology to automatize digital radiography defect inspection. Traditionally, in an industrial application, conventional methods are inefficient when the detection targets are small, local, and subtle in the complex scenario. Meanwhile, the outperformance of deep learning models, such as the convolutional neural network (CNN), is limited by a huge volume of data with precise annotations. To overcome these challenges, an efficient CNN model, only trained with image-level labels, is first proposed for detection of tiny casting defects in a complicated industrial scene. Then, in this article, we present a novel training strategy which can form a new object-level attention mechanism for the model during the training phase, and bilinear pooling is utilized to improve the model capability of detecting local contrast casting defects. Moreover, to enhance the interpretability, we extend class activation maps (CAM) to bilinear CAM (Bi-CAM) which is adapted to bilinear architectures as a visualization technique to reason about the model output. Experimental results show that the proposed model achieves superior performance in terms of each quantitative metric and is suitable for most actual applications. The real-time defect detection of castings is efficiently implemented in the complex scenario.**

***Index Terms*—Bilinear class activation maps (Bi-CAM), bilinear pooling, convolutional neural network (CNN), digital radiography (DR) defect inspection, object-level attention mechanism.**

## I. Introduction

ALUMINIUM castings have been extensively applied to various parts of automobile [1], aerospace [2], aircraft [3], and electric devices [4] whose qualities affect the fatigue behavior of the overall product [5]. Due to the complexity and diversity of the casting process [6], defects are inevitable during the process, such as gas holes, sand holes, and flaws. And these defects cannot be recognized by surface detection technology when the defects exist in the interior of castings. In order to obtain the internal in formation and guarantee the completeness of castings, radiography is often used for nondestructive testing [7], which has been widely used in quality controlling [8] and security inspection system [9]. In radiographic testing processing, most manufacturers rely mainly on manual detection according to the experiences of operators about the shape, brightness, and contrast of castings. Such a manual method is not only mechanical and inefficient, but also may cause ophthalmic diseases due to the high-frequency illumination of the display. Consequently, the digital radiography (DR)-based automatic inspection system has been one of the research focuses.

Over the past decades, conventional computer vision technology has extensive applications in many automatic inspection systems [10]–[13]. Handcrafted features and statistic-based machine learning models are often the main tools, which use image algorithms to generate the feature vectors of texture, color, shape, and spectral cues, and then, adopt statistical machine learning models to realize inspection systems. With the advantage of simply and effective models, machine learning methods achieve satisfactory performances in most applications. However, the general drawback of conventional methods is that their performances depend on the effective representation extracted by handcrafted feature algorithms. Designing a reliable and robust representation may not be efficient and requires expensive expert knowledge. Moreover, the common character in these previous application scenarios is that the detection background and the structure of targeted objects are relatively single and salient. In our scenario, the position and structure of castings may be various, and the defects on our radiography images are subtle and local, as illustrated in Fig. 1. Thus, it is difficult to obtain the effective feature representation by conventional handcrafted feature algorithms, and especially so in the complicated scenario with the various casting structure, free position, and small local defects.

Recently, deep learning model, particularly the convolutional neural network (CNN), has received substantial interest in industrial applications [14]–[17]. CNNs are capable of automatically extracting more discriminative and hierarchical features,

Manuscript received August 6, 2019; revised November 1, 2019; accepted December 12, 2019. Date of publication January 1, 2020; date of current version August 18, 2020. This work was supported in part by the National Natural Science Foundation of China under Grant 61673276. *(Chuanfei Hu and Yongxiong Wang contributed equally to this work.) (Corresponding author: Yongxiong Wang.)*

The authors are with the Department of Control Science and Engineering, University of Shanghai for Science and Technology, Shanghai 200093, China (e-mail: w64228013@126.com; wyxiong@usst.edu.cn).

Color versions of one or more of the figures in this article are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TIE.2019.2962437

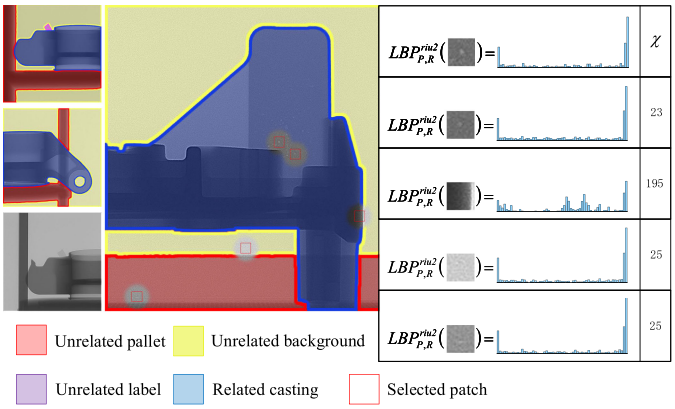



Fig. 1. On the left, the various regions of the radiography images are masked by different colors. The result reveals the complicacy of the scenario with various structures and positions of castings. Besides, uniform local binary pattern (LBP) algorithm is introduced which can generate a histogram by coding each pixel from a region into a binary pattern according to adjacent pixels. We extract uniform LBP features of the image patches which are selected from related or unrelated areas. On the right, the chi-square distances of the LBP histograms are computed for the defect patch (in the first row of the table) to the others. The comparison shows that the features of defects appear in the other local regions. The locality attribute of casting defects is revealed. Uniform LBP algorithm is denoted as $\mathrm{LBP}^{\mathrm{riu2}}_{P,R}(\cdot)$ where $P$ represents the number of local neighbors ($P = 8$) on a circle of radius $R$ ($R = 2$).

which can save huge labors of designing complicated feature algorithms. In DR defect inspection, which requires quick and reliable results, the powerful deep learning model may not be the best solution due to its huge computation and high dependence of precise annotations. The performance of the deep learning model generally improves with the growth of model parameters, which would unavoidably raise the computing consumptions. It is a main issue to make a good tradeoff between the performance and efficiency of execution in a real-time system. For another problem, labeling each defect in each radiography image is very expensive and laborsome. An alternative way is to annotate the images with coarse labels only indicating whether there are any defects in the image or not. However, an image-level label does not have enough position information of the defects. It is not easy for the classification model to extract defect features directly, because the differences between defect and nondefect images include not only defects, but also the complex structures of castings and background in the complicated scenario.

To overcome the challenges of subtle defect representations and lack of precise annotations, we introduce attention mechanism, bilinear pooling, and depthwise separable convolution to the defect detection model. Attention mechanism, which is able to select more interested attributes of input data, has been proposed and successfully used in many different tasks [18]–[21]. For vision tasks, attention mechanism enables the model to focus more on the interested regions, and ignore the others that contribute less to the final results. In this way, the attention mechanism can be leveraged to reduce the disturbance of unrelated regions and fill the blank of missing position information when the defect detection model is trained with coarse labels. Different from other patterns of attention mechanism, we propose a novel training strategy to form a new object-level attention mechanism which can achieve identical effect without extra structures. The obvious advantage is that the model avoids additional computing burdens. With the success of bilinear pooling in fine-grained image classification tasks [22], the bilinear pooling has been utilized as a common method to obtain deeper and superior feature representation [23]–[25]. This aggregation method not only retains the second-order original features but also mines more correlation information between each feature element. Therefore, we leverage the bilinear pooling to represent subtle defect features, which is attractive and feasible. Moreover, in order to achieve a real-time model, depthwise separable convolution [26] is introduced to our model. In summary, the main contributions are as follows.

1) A new CNN model, including type classification module (TCM) and defect classification module (DCM), is constructed for DR defect inspection, which is capable of mining object-related features. Simultaneously, bilinear pooling is used to obtain strong feature representation, and depthwise separable convolution is introduced in DCM to reduce computation. In real-time complex application, the proposed model outperforms classical deep classification models in terms of each quantitative metric (e.g., accuracy, precision, recall, F-measure, FPS, and Para Size).
2) Compared with previous methods of attention mechanism, a novel training strategy is proposed to construct object-level attention mechanism which is effective and free of computing burdens. To the best our knowledge, we are among the first to achieve the attention mechanism without additional architectures for defect detection models.
3) Inspired by class activation maps (CAM) [27], we propose bilinear CAM (Bi-CAM), which is suitable to bilinear architectures, to be used as a visualization technique to increase the interpretability of the model.

The remainder of this article is organized as follows. Section II presents an overview of recent studies for automatic defect detection methods and attention mechanism. Section III articulates the main method in details. Section IV includes the experimental results and analysis. Finally, Section V concludes this article.

## II. Related Works

### A. Automatic Defect Detection Methods

In the field of industrial inspection, many machine learning solutions have been proposed to process and analyze industrial images for detecting defects. These conventional methods can be generally divided into computing based and machine learning based. In computing-based methods, numerical information of images, such as color [28]–[30] and frequency [31], [32], is utilized to segment the defect regions from original images. The advantage of the computing-based method is that it is concise and easy to be comprehended visually. However, the selection of threshold is an unavoidable difficulty due to the susceptibility of noise and clutter. Machine learning-based methods have been proposed more often, which utilize feature algorithms to

construct sample set and execute the detection inference from the statistical aspect [12], [13]. They can avert the repeated threshold testing and simplify the generalization of models.

Recently, deep learning has become a trend in the development of machine learning and achieved great success in computer vision. CNN is one of the most successful deep models which can extract discriminate features by many convolutional operators. To automatically extract features and achieve end-to-end methods, many researchers and designers gradually propose the inspection methods based on CNN and its varieties. A wide-and-compact CNN architecture is constructed by Li *et al.* [33] that can learn multiscale and multiposition features from fabric, and achieve the automatic detection of fabric defects. Cheon *et al.* [34] design a CNN to classify defects of wafer surface and utilize feature vectors extracted by the network to construct a $k$-NN model for classification of unknown defect class. Haselmann *et al.* [35] construct an autoencoder by CNN and inject random artificial defects into samples to train the network to detect locally confined defects on fabric surface without manual labels. Li *et al.* [36] develop a new generative model and mean-covariance labeling scheme to expand the data space. This can be seen as preprocessing for the reliable classification of gear safety to improve the performance of general classifiers. With the review of previous works, the application scenario of existing methods is simple and the detected object is regular. However, they may lose efficacy when the detection environment becomes various and complex. Thus, certain effective module or mechanism may bring us new insight to overcome the impact of clutter scenario, such as attention mechanism.

### B. Attention Mechanism

In recent years, inspired by human perception process [37], attention mechanism has extensive applications [18]–[21]. As our detection method is based on vision, we mainly focus on current studies of visual attention mechanism. Xu *et al.* [38] propose an attention-based model for image caption tasks. "Soft" deterministic attention and "hard" stochastic attention are introduced in the model where they are trained by different propagation algorithms. Hu *et al.* [39] propose a squeeze-and-excitation module which utilizes global average pooling to abstract the internal features, and computes the weights for each feature by linear combination and nonlinear activation. This module can be inserted into the general networks where a channel-level attention is achieved to enhance the representation of features. Based on previous works, Woo *et al.* [40] propose convolutional block attention module (CBAM), which is composed of channel-level attention module and spatial-level attention module. CBAM can selectively learn features by channels and areas. Although these attention methods can improve the performance of models for most vision tasks, they may not be suitable to industrial inspection. Furthermore, most of these methods should inevitably increase the structure of networks and the additional computation which is undesirable in a real-time application. To construct a light and effective attention mechanism is still an open issue for practical applications.

## III. Proposed Method

In this section, we first explain the proposed training strategy to generate a new object-level attention mechanism without additional computing burdens. Then, the details of two modules, TCM and DCM, are discussed. The proposed visualization technique Bi-CAM, which is designed for the bilinear pooling structure, is described in Section III-D.

### A. Training Strategy for Object-Level Attention Mechanism

In our task, any spatial information of defects cannot be provided because the dataset is only annotated with image-level labels, defect or nondefect. Thus, attention mechanism is an indispensable tool in complicated scenarios. The concept of attention mechanism derives from human brain instinct that odd objects are attractive for the cerebral cortex in the field of vision [41]. For deep learning-based vision tasks, previous works construct many kinds of additional modules to simulate this mechanism, and successfully improve the performance. Furthermore, some special modules abstract the attention concept from spatial level to channel level [42]. Although these modules are flexible to insert standard networks, the model size and computation would increase inevitably.

In order to guarantee an industrial inspection in real time, we propose a novel training strategy without additional network structures to form an attention mechanism, namely, object-level attention mechanism. The basic idea of the scheme is motivated by the process of teaching infants, where parents often teach infants to pay more attention to recognizing the low-level objects first, such as fruits and bottles, then infants are told whether they can eat these objects. Inspired by this way, the proposed training strategy is to let the model pay attention to the certain type of the casting in the image, and then, the model can be taught to infer defects accurately based on the cognition. In the first stage of the proposed strategy, the object-level attention mechanism which bridges the gap between image-level annotations and application scenario is generated. Different from transfer learning where the existing knowledge is learned to solve a new but related task [43], the core perspective of our training strategy is to train the model gradually from shallow task to deep task, rather than a simple knowledge transfer.

According to the proposed strategy, we set two datasets with type and defect labels, respectively. Simultaneously, a novel CNN model is constructed where the two subnetworks are cascaded, namely, TCM and DCM. During the strategy, we first train TCM on the type dataset with an additional classifier to distinguish the type of castings, then the softmax layer is discarded after TCM converges. The trained TCM can extract more object-related features of the casting in the radiography image. In secondary stage, the total model is trained on the defect dataset while the parameters of TCM are fixed. That means that only DCM are trained. The object-level attention mechanism enables TCM to provide object-related features for subsequent DCM. Intuitively, DCM is able to learn the defect features from effective areas with image-level supervision. Overall framework of the method is depicted in Fig. 2.

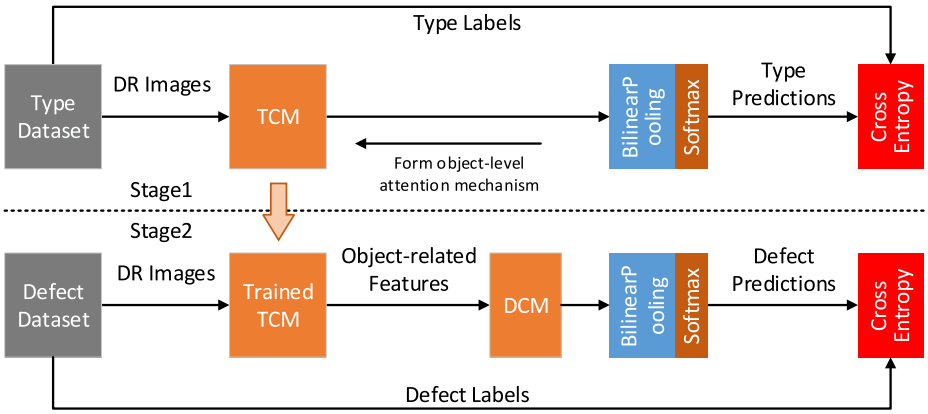



Fig. 2. Overall framework of the proposed method.

## B. Type Classification Module

TCM is in charge of irrelevant suppression and object-related representation in complex scenario via the object-level attention mechanism. We implement the TCM based on VGG16 [44] which is effective and light. Given an input image, the feature maps are extracted by TCM which can be formulated as

$$f_{\mathrm{TCM}}\left(X; w_{\mathrm{TCM}}, b_{\mathrm{TCM}}\right) = w_{\mathrm{TCM}} * X + b_{\mathrm{TCM}} \tag{1}$$

where $X$ is the input radiography image, $w_{\mathrm{TCM}}$ and $b_{\mathrm{TCM}}$ are the parameters of convolutional layers in TCM which can be learned, $*$ represents the convolutional operation. $f_{\mathrm{TCM}}(X; w_{\mathrm{TCM}}, b_{\mathrm{TCM}})$ is the output feature maps of TCM. Typically, nonlinear action functions (ReLU [45]) and pooling layers follow convolutional layers in order to improve translation invariance and representation of features. To simplify the expression, these are omitted above the formulation.

We notice that the downsampling factor of output features in these original architectures is particularly critical. For example, when the input of the network is a $448 \times 448$ image, you can only obtain $14 \times 14$ feature which is too small to contain enough spatial information of the casting. Meanwhile, in order to demonstrate the universality, matured network structures are introduced to our method. We also obverse the diminishing returns in preliminary comparative trials that occurred by increasing parameters simply, such as the number or sizes of layers. Thus, the structures are merely modified to guarantee the size of features relatively, as listed in Table I. We retain the original structure of convolutional layers and discard the last max pooling layer. Deeper features, which contain more discriminative information, are helpful to improve the final distinction. This is the motivation that we only adopt the last portion of TCM.

As mentioned in Section I, to improve the representation of subtle defects, the bilinear pooling is inserted before prediction layer. Assuming that the output of the module is $X \in R^{L \times D}$, $Y = X$, where $L$ and $D$ are the number and channel of features, respectively. Then, the bilinear aggregation is denoted as

$$Z = X^T Y \tag{2}$$

TABLE I
ARCHITECTURE OF TCM

| Layer | Kernel Shape | Kernel Number | Stride | Output |
|---|---|---|---|---|
| Conv | $3 \times 3$ | 64 | 1 | |
| Conv | $3 \times 3$ | 64 | 1 | |
| Pool | $2 \times 2$ | 1 | 2 | $224 \times 224$ |
| Conv | $3 \times 3$ | 128 | 1 | |
| Conv | $3 \times 3$ | 128 | 1 | |
| Pool | $2 \times 2$ | 1 | 2 | $112 \times 112$ |
| Conv | $3 \times 3$ | 256 | 1 | |
| Conv | $3 \times 3$ | 256 | 1 | |
| Conv | $3 \times 3$ | 256 | 1 | |
| Pool | $2 \times 2$ | 1 | 2 | $56 \times 56$ |
| Conv | $3 \times 3$ | 512 | 1 | |
| Conv | $3 \times 3$ | 512 | 1 | |
| Conv | $3 \times 3$ | 512 | 1 | |
| Pool | $2 \times 2$ | 1 | 2 | $28 \times 28$ |
| Conv | $3 \times 3$ | 512 | 1 | |
| Conv | $3 \times 3$ | 512 | 1 | |
| Conv | $3 \times 3$ | 512 | 1 | $28 \times 28$ |
| Bilinear | | | | |
| Softmax | | | | |

where $Z$ is the result of matrix multiplication for $X$ and $Y$. Following [22], signed square root step and $\ell_2$ normalization are inspired to be applied for bilinear pooling result $Z$. Overall bilinear aggregation can be integrated into an end-to-end network, because its gradient computation is available.

The output nodes of the last layer are the same as the number of type classes. The softmax activation function can be regarded as an estimation of probability of each class since the sum of softmax classifier is one. We denote the probability of class $i$ as $p_i$ which can be calculated as follows:

$$p_i = \frac{\exp\left(w_i x + b_i\right)}{\sum_J \exp\left(w_j x + b_j\right)} \tag{3}$$

where $x$ is the input vector of softmax classifier belonging to class $i$, $w_i$ and $b_i$ refer to the weight vector and bias of output

TABLE II
ARCHITECTURE OF DCM

| | Layer | Kernel Shape | Kernel Number | Stride | Output |
|---|---|---|---|---|---|
| ×3 | DConv | 1 × 1 | 512 | 1 | 28 × 28 |
| | BN | | 512 | | |
| | Conv | 1 × 1 | 512 | 1 | |
| | BN | | 512 | | |
| | DConv | 3 × 3 | 512 | 1 | |
| | BN | | 512 | | |
| | Conv | 1 × 1 | 512 | 1 | |
| | BN | | 512 | | |
| | DConv | 1 × 1 | 2048 | 1 | |
| | BN | | 2048 | | |
| | Conv | 1 × 1 | 2048 | 1 | |
| | BN | | 2048 | | 28 × 28 |
| | Shortcut | | | | |
| | Bilinear | | | | |
| | Softmax | | | | |

node $i$, respectively. The number of classes is denoted by $J$. We consider the class with the highest probability as the final result from the classifier. In the training stage, the cross-entropy loss function is defined as

$$\mathcal{L}_{\text{cross}} = -\sum_{J} y_j \ln p_j \tag{4}$$

where $\mathcal{L}_{\text{cross}}$ is leveraged to measure the error between the result $p_j$ and the target label $y_j$.

### C. Defect Classification Module

TCM and DCM are quietly correlated but essentially different. TCM leverages an object-level attention mechanism to suppress the impact of unrelated regions. DCM aims at achieving a tiny defect classification by mining deeper features in the object-related features from TCM. To meet industrial inspection requirements of accuracy, speed, and size of model, we attempt to design DCM by stacking few computational units which are efficient and similar with residual blocks. Due to the efficiency of depthwise separable convolution [26], we replace entire standard convolutions to the depthwise separable convolutions in residual blocks. The data stream of DCM can be formulated as

$$f_{\text{DCM}}\left(X; w_{\text{DCM}}, b_{\text{DCM}}\right) = w_{\text{DCM}} \otimes f_{\text{TCM}}(X) + b_{\text{DCM}} \tag{5}$$

where $f_{\text{TCM}}(X)$ are the feature maps from TCM, $w_{\text{DCM}}$ and $b_{\text{DCM}}$ are learnable parameters of depthwise separable convolutional layers in DCM. $\otimes$ represents the depthwise separable convolutional operation. As listed in Table II, DCM is composed of three stacked residual units constructed by the depthwise separable convolutions. Before the softmax layer, the bilinear pooling is also used to enhance the representation of the subtle defects. We still use cross-entropy loss to train DCM in secondary training stage.

### D. Bilinear Class Activation Maps

In order to interpret the validity of the proposed model visibly, we require a visual method to depict the relationship between feature maps and predictions. Inspired by CAM [27], we propose a new method, called Bi-CAM, to generate the active feature map for bilinear architectures by the simple algorithm. The procedure for generating CAM $M_c(x, y)$ of class $c$ is formulated by

$$M_c(x, y) = \sum_k w_k^c f_k(x, y) \tag{6}$$

where $f_k(x, y)$ represents the feature map of channel $k$ in the last convolutional layer of the classification network at spatial $(x, y)$. $w_k^c$ is the weight corresponding to class $c$ in the softmax layer for channel $k$.

However, the original procedure is only suitable to global average pooling (GAP) architectures. To adapt bilinear architectures in our model, we extend the CAM method. The bilinear pooling is instanced such that

$$Z = X^T Y = X^T X = \begin{pmatrix} f_1^T f_1 & \cdots & f_1^T f_D \\ \vdots & \ddots & \cdots \\ f_D^T f_1 & \cdots & f_D^T f_D \end{pmatrix} \tag{7}$$

where $Z$ is a Gram matrix associated with feature map $f_k$, $k = 1, \ldots, D$, where $f_k$ is an abbreviation of $f_k(x, y)$, $z_{ij} = f_i^T f_j$ is the outer product of $f_i$ and $f_j$, $i, j \in k$. Following [27], but being quite different, the score of class $c$ in the softmax layer is denoted by

$$S_c = \sum_{i,j \in k} w_{ij}^c f_i^T f_j = \sum_{i,j \in k} w_{ij}^c z_{ij} \tag{8}$$

where $w_{ij}^c$ is the weight corresponding to $z_{ij}$.

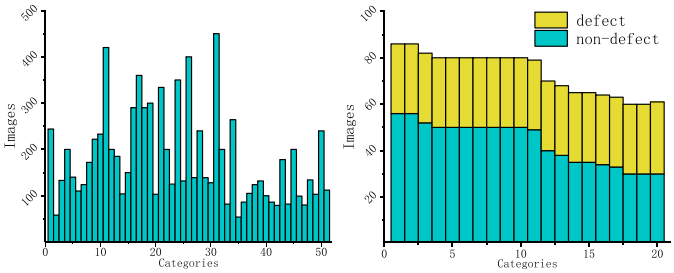
To spotlight the active regions of predicted class, a set of weights is required to describe the activation of each feature map. Thus, we first construct a square matrix consisting of $w_{ij}^c$. Second, eigenvalues $\widetilde{w}_{ii}^c$ are obtained by eigen-decomposition approach. $\widetilde{w}_{ii}^c$ can approximate the importance of $f_i$, since more information of each channel is concentrated in the corresponding eigenvalue. Finally, Bi-CAM is formulated by

$$M_c^{bi} = \sum_{i \in k} \widetilde{w}_{ii}^c f_i. \tag{9}$$

The effect of Bi-CAM is similar to CAM, in the sense that Bi-CAM is a novel derivation of CAM for bilinear architectures. $M_c^{bi}$ is only an initial heat-map of the same size as the last convolutional feature maps. By upsampling $M_c^{bi}$ to the size of the input image, image regions can be highlighted which are most relevant to the particular class.

## IV. EXPERIMENT

In this section, the experimental setup, which includes the used datasets, the implementation details, and the evaluation metrics, is described first. Then, we present the quantitative comparisons with classical models on the defect dataset. Finally, a series of ablation studies investigate the impact of the proposed method in two modules. To understand and show the improvement directly, the visualization technique is used as a powerful tool to support experimental results.




(a) type dataset (b) defect dataset

Fig. 3. (a) and (b) Image numbers of each category in the type and defect dataset. Compared with the distribution of the type dataset in (a), the impact of unbalanced samples would be scaled while the samples are not abundant. Thus, the defect dataset is constructed carefully to guarantee the balance between defect and nondefect samples in (b). Yellow and green bars represent the numbers of defect and nondefect images, respectively.

## A. Experimental Setup

***1) Datasets:*** The radiography images, provided by the cooperative enterprise, are generated by the inspection equipment named Y.MU2000-D which is an industrial X-ray and CT inspection system produced by YXLON. The original samples are exported by CT analysis station and the sizes are fixed to $1000 \times 1000$. In order to adapt our proposed training strategy, the dataset is divided into type and defect datasets which are independent, respectively. Concurrently, the balance between each category in the datasets, especially small-scale dataset, is also considered to avoid the impact of unbalanced samples during our experiments. Thus, compared with the type dataset, the sample distribution of the defect dataset is considered more cautiously. The type dataset includes 9215 images of 51 categories for each type of castings, as shown in Fig. 3(a). TCM is trained on the training set which randomly selected 70% images from each category of the type dataset. We evaluate the module on the testing set on the remaining images. As defect images produced by the system are rare, the defect dataset includes 1469 images of 20 categories only with image-level defect labels, as shown in Fig. 3(b). We alter the distribution of training set and testing set to 60% and 40% in the defect dataset in order to verify the generalization ability of the model. Moreover, our model is tested on the defect dataset for five times where the training set and testing set are repartitioned in each trial.

***2) Implementation Details:*** The experiments are conducted on a work station with an Intel Xeon Silver 4116 CPU and a TITAN XP GPU. The proposed model is based on a deep learning framework PyTorch. As mentioned in Section III, TCM and DCM are trained in two stages. The same training details are: 1) initial parameters pretrained on the ImageNet [46]; 2) training with 40 epochs; and 3) optimizer which uses stochastic gradient descent (SGD) with 0.9 momentum parameter and 0.0005 weight decay. The differences are the policies of learning rate and mini-batch. In first stage, we fix the learning rate at 0.01 and the mini-batch is set to 16. In the secondary stage, the step policy is utilized to divide the initial learning rate 0.01 in 10 by 15 epochs each, the mini-batch is set to 4. In the experiments, all input radiography images are extended to three channels and resized to $448 \times 448$.

TABLE III
COMPARISONS WITH VGG16 AND RESNET50

| Model | Accuracy(%)↑ | Precision(%)↑ | Recall(%)↑ | F-M(%)↑ | FPS↑ | Para Size(MB)↓ |
|---|---|---|---|---|---|---|
| VGG16 | 54.51 | 45.51 | 57.08 | 50.65 | 40.67 | 1688.19 |
| ResNet50 | 56.90 | 46.52 | 36.25 | 40.75 | 41.07 | 89.99 |
| Our model | 92.77 | 90.67 | 91.60 | 91.31 | 49.11 | 58.27 |

***3) Evaluation Metrics:*** We adopt accuracy, precision, recall, F-Measure (F-M), frames per second (FPS), and parameter size (Para Size) as the metrics to evaluate our method comprehensively. The aforesaid metrics are defined as

$$\text{Accuracy} = T/S \tag{10}$$

$$\text{Precision} = \text{TP}/(\text{TP+FP}) \tag{11}$$

$$\text{Recall} = \text{TP}/(\text{TP+FN}) \tag{12}$$

$$\text{F-Measure} = \frac{2 * \text{Recall} * \text{Precision}}{\text{Recall} + \text{Precision}} \tag{13}$$

where $T$ denotes to the number of successfully classified images with corresponding classes, and $S$ denotes the number of test images. TP and FN are the numbers of defect images which are detected correctly or not. FP represents the number of nondefect images which are misclassified. In our experiment, the results of FPS are the average of 500 testing images.

## B. Comparisons With Other Classical Models

Our model is compared with VGG16 and ResNet50, and testing performances on our defect dataset are presented in Table III. The results are as expected and show that VGG16 and ResNet50 are far away from the proposed model on each quantitative evaluation metric, such as accuracy, precision, recall, F-Measure, FPS, and Para Size. We also try to adjust the training hyperparameters in these classical models. However, they do not exhibit the convergence on our defect dataset. The reason for the low performances may be that the standard models lack the capability of distinguishing subtle differences in the complicated background. For our model, object-level attention mechanism can suppress the disturbed regions and replenish rough position cues. Concurrently, bilinear pooling and depthwise separable convolutions assist the model to construct stronger feature representation for subtle defects effectively. The results agree with the proposed improvements which bridge the gap between the trial and application. We further test their effectiveness by running a series of ablation experiments and showing the corresponding quantitative and visual results in the next subsection.

## C. Ablation Studies

***1) Effect of Bilinear Pooling in TCM:*** To clarify the effect of bilinear pooling in TCM, two different structures are designed based on bilinear pooling and global average pooling. In Table IV, it has been shown that TCM based on bilinear pooling possesses good performance. The relative improvements of each metric approximates 3% to 7%. We can consider that TCM is

TABLE IV
CLASSIFICATION RESULTS OF THE PROPOSED MODEL ON THE DEFECT DATASET

| -gap | -bi | Accuracy(%)↑ | Precision(%)↑ | Recall(%)↑ | F-M(%)↑ |
|---|---|---|---|---|---|
| ✓ | | 89.16 | 87.84 | 85.40 | 86.67 |
| | ✓ | 92.43 | 90.14 | 91.20 | 90.83 |
| Relative improvement (%) | | 3.668 | 2.618 | 6.792 | 4.799 |

The model is based on different architectures of TCM. Global average pooling and bilinear pooling are denoted by "-gap" and "-bi," respectively.

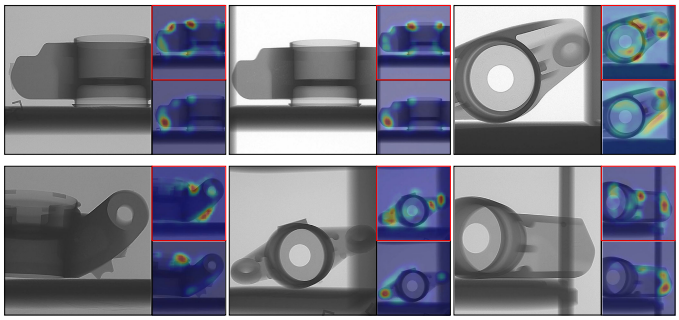


Fig. 4. Some examples of CAM and Bi-CAM, where the images based on Bi-CAM are labeled with the red rectangle.

TABLE V
CLASSIFICATION RESULTS OF THE PROPOSED MODEL ON THE DEFECT DATASET WITH THE COMBINATIONS OF DIFFERENT COMPONENTS

| Combination | # | Accuracy(%)↑ | Precision(%)↑ | Recall(%)↑ | F-M(%)↑ | FPS↑ | Para Size(MB)↓ |
|---|---|---|---|---|---|---|---|
| TCM-gap+DCM-gap | | 58.57 | - | - | - | 39.42 | 140.44 |
| TCM-bi+DCM-gap | | 59.18 | - | - | - | 39.42 | 140.44 |
| TCM-bi+DCM-gap | ✓ | 92.43 | 90.14 | 91.20 | 90.83 | 39.42 | 140.44 |
| TCM-bi+DCM-gap-ds | ✓ | 91.92 | 90.11 | 89.60 | 90.12 | 53.79 | 56.32 |
| TCM-bi+DCM-bi | ✓ | 93.49 | 92.58 | 92.80 | 92.99 | 24.16 | 172.42 |
| TCM-bi+DCM-bi-ds | ✓ | 92.77 | 90.67 | 91.60 | 91.31 | 49.11 | 58.27 |

We use "-bi" and "-ds" to denote a module based on bilinear pooling and depthwise separable convolution, respectively. "–" is a substitute for very poor results. "#" means the model is trained with the proposed strategy.

guided to provide much subtle object-related features for DCM. Meanwhile, Bi-CAMs and CAMs are depicted for the identical categories in Fig. 4. As can be seen, Bi-CAM surely highlights more details of castings than CAM. This is a visualization interpretation for the effect of bilinear pooling in TCM.

***2) Effect of Object-Level Attention Mechanism:*** To confirm the proposed training strategy which can generate an object-level attention mechanism, our model is trained twice by the proposed strategy and conventional strategy, respectively. From the first three rows of Table V, it can be demonstrated that the proposed strategy plays the decisive role in the model convergence without additional computational burdens.

***3) Contribution of Depthwise Separable Convolution and Bilinear Pooling to DCM:*** The effect of depthwise separable convolutions can be seen in Table V. By comparing the DCM-gap-ds with DCM-gap, we find that depthwise separable convolution dramatically reduces the parameter size of the overall model and improves the FPS by 36.45%. Meanwhile, the effect of bilinear operation is investigated for DCM and

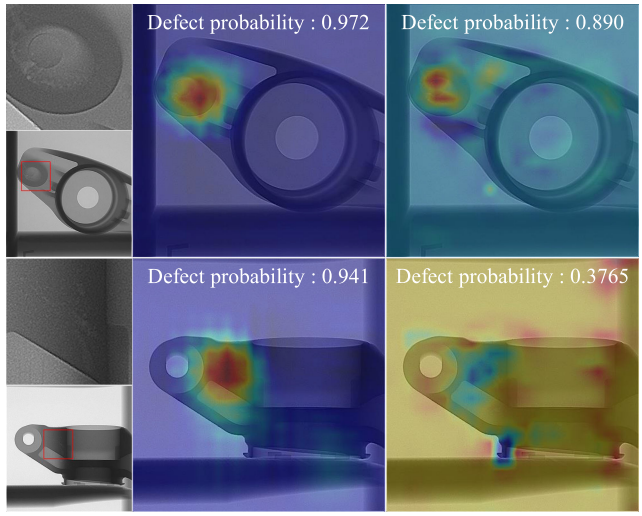



Fig. 5. Examples of Bi-CAM and CAM, in which the first column contains original images and defect regions, the second to third columns are Bi-CAMs and CAMs. The prediction probability of defects is shown in each image.

DCM-ds. In Table V, the bilinear pooling achieves an improvement of about 2% in the metrics of accuracy and recall. We also illustrate activation maps of DCM based on "-bi" and "-gap" in Fig. 5. For the first row of the comparisons, defect regions can be highlighted more precisely by Bi-CAM. It can be considered as a visual explanation for the higher prediction probability of the defects. Similarly, the false prediction is shown in the second example. The model based on "-gap" is hard to focus on local contrast defect regions, thus more unrelated regions are responded. These results demonstrate that bilinear pooling can make the proposed model practical to detect tiny defects on radiography images.

## V. CONCLUSION

In this article, we proposed a novel training strategy to form a new object-level attention mechanism, and construct an efficient CNN model based on the object-level attention mechanism and bilinear pooling for the DR defect inspection. The defects of the casting on radiography images were effectively recognized in complicated detection scenario. The proposed model outperformed other classical deep learning classification models in each quantitative metric. It demonstrated that our model has a reliable advantage in efficiency and can be used for a real-time DR inspection system with complicated background.

Researchers should make the best use of training strategies or network architectures to improve the performance and reduce additional labor consumption, such as manual annotation. This is a trend of designing industrial detection model. However, compared to direct supervision methods, the drawback is that our model cannot predict excessive results, such as the positions and types of defects. In future work, we aim to investigate topology analysis to infer a coarse location by the correlation of features and utilize metric learning to divide the types of defects by the learnable model.

## REFERENCES


[1] W. S. Miller *et al.*, "Recent development in aluminium alloys for the automotive industry," *Mater. Sci. Eng. A*, vol. 280, no. 1, pp. 37–49, Mar. 2000.

[2] A. Heinz, A. Haszler, C. Keidel, S. Moldenhauer, R. Benedictus, and W. S. Miller, "Recent development in aluminium alloys for aerospace applications," *Mater. Sci. Eng. A*, vol. 280, no. 1, pp. 102–107, Mar. 2000.

[3] T. Dursun and C. Soutis, "Recent developments in advanced aircraft aluminium alloys," *Mater. Des. (1980–2015)*, vol. 56, no. 4, pp. 862–871, Apr. 2014.

[4] M. Riera-Guasp, J. A. Antonino-Daviu, and G. Capolino, "Advances in electrical machine, power electronic, and drive condition monitoring and fault detection: State of the art," *IEEE Trans. Ind. Electron.*, vol. 62, no. 3, pp. 1746–1759, Mar. 2015.

[5] H. Mayer, "Recent developments in ultrasonic fatigue," *Fatigue Fracture Eng. Mater. Struct.*, vol. 39, no. 1, pp. 3–29, Jan. 2016.

[6] S. Pattnaik, D. B. Karunakar, and P. K. Jha, "Developments in investment casting process—A review," *J. Mater. Process. Technol.*, vol. 212, no. 11, pp. 2332–2348, Nov. 2012.

[7] Y. Hangai *et al.*, "Nondestructive observation of pore structure deformation behavior of functionally graded aluminum foam by X-ray computed tomography," *Mater. Sci. Eng., A*, vol. 556, pp. 678–684, Oct. 2012.

[8] J. Wang, P. Fu, and R. X. Gao, "Machine vision intelligence for product defect inspection based on deep learning and hough transform," *J. Manuf. Syst.*, vol. 51, pp. 52–60, 2019.

[9] D. Mery, E. Svec, M. Arias, V. Riffo, J. M. Saavedra, and S. Banerjee, "Modern computer vision techniques for X-ray testing in baggage inspection," *IEEE Trans. Syst., Man Cybern., Syst.*, vol. 47, no. 4, pp. 682–692, Apr. 2017.

[10] D. You, X. Gao, and S. Katayama, "Multisensor fusion system for monitoring high-power disk laser welding using support vector machine," *IEEE Trans. Ind. Informat.*, vol. 10, no. 2, pp. 1285–1295, May 2014.

[11] R. Shanmugamani, M. Sadique, and B. Ramamoorthy, "Detection and classification of surface defects of gun barrels using computer vision and machine learning," *Measurement*, vol. 60, pp. 222–230, Jan. 2015.

[12] J. Tian, C. Morillo, M. H. Azarian, and M. Pecht, "Motor bearing fault detection using spectral kurtosis-based feature extraction coupled with k-nearest neighbor distance analysis," *IEEE Trans. Ind. Electron.*, vol. 63, no. 3, pp. 1793–1803, Mar. 2016.

[13] N. Boaretto and T. M. Centeno, "Automated detection of welding defects in pipelines from radiographic images DWDI," *NDT E Int.*, vol. 86, pp. 7–13, Mar. 2017.

[14] M. S. Hossain, M. H. Al-Hammadi, and G. Muhammad, "Automatic fruits classification using deep learning for industrial applications," *IEEE Trans. Ind. Informat.*, vol. 15, no. 2, pp. 1027–1034, Feb. 2019.

[15] Q. Xuan, Z. Chen, Y. Liu, H. Huang, G. Bao, and D. Zhang, "Multiview generative adversarial network and its application in pearl classification," *IEEE Trans. Ind. Electron.*, vol. 66, no. 10, pp. 8244–8252, Oct. 2019.

[16] W. Chen, D. Ding, J. Mao, H. Liu, and N. Hou, "Dynamical performance analysis of communication-embedded neural networks: A survey," *Neurocomputing*, vol. 346, pp. 3–11, Jun. 2019.

[17] H. Yang, Y. Chen, K. Song, and Z. Yin, "Multiscale feature-clustering-based fully convolutional autoencoder for fast accurate visual inspection of texture surface defects," *IEEE Trans. Autom. Sci. Eng.*, vol. 16, no. 3, pp. 1450–1467, Jul. 2019.

[18] W. Li, X. Zhu, and S. Gong, "Harmonious attention network for person re-identification," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 2285–2294.

[19] H. Choi, K. Cho, and Y. Bengio, "Fine-grained attention mechanism for neural machine translation," *Neurocomputing*, vol. 284, pp. 171–176, Apr. 2018.

[20] J. Liu, G. Wang, L. Duan, K. Abdiyeva, and A. C. Kot, "Skeleton-based human action recognition with global context-aware attention LSTM networks," *IEEE Trans. Image Process.*, vol. 27, no. 4, pp. 1586–1599, Apr. 2018.

[21] H. Chen and Y. Li, "Three-stream attention-aware network for RGB-D salient object detection," *IEEE Trans. Image Process.*, vol. 28, no. 6, pp. 2825–2835, Jun. 2019.

[22] T. Lin, A. RoyChowdhury, and S. Maji, "Bilinear convolutional neural networks for fine-grained visual recognition," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 40, no. 6, pp. 1309–1322, Jun. 2018.

[23] Z. Yu, J. Yu, J. Fan, and D. Tao, "Multi-modal factorized bilinear pooling with co-attention learning for visual question answering," in *Proc. IEEE Int. Conf. Comput. Vis.*, Oct. 2017, pp. 1839–1848.

[24] A. Alzu'bi, A. Amira, and N. Ramzan, "Content-based image retrieval with compact deep convolutional features," *Neurocomputing*, vol. 249, pp. 95–105, Aug. 2017.

[25] L. Wu, Y. Wang, X. Li, and J. Gao, "Deep attention-based spatially recursive networks for fine-grained visual recognition," *IEEE Trans. Cybern.*, vol. 49, no. 5, pp. 1791–1802, May 2019.

[26] F. Chollet, "Xception: Deep learning with depthwise separable convolutions," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jul. 2017, pp. 1800–1807.

[27] B. Zhou, A. Khosla, A. Lapedriza, A. Oliva, and A. Torralba, "Learning deep features for discriminative localization," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2016, pp. 2921–2929.

[28] M. Aminzadeh and T. Kurfess, "Automatic thresholding for defect detection by background histogram mode extents," *J. Manuf. Syst.*, vol. 37, no. 1, pp. 83–92, Oct. 2015.

[29] L. Bai, X. Yang, and H. Gao, "A novel coarse–fine method for ball grid array component positioning and defect inspection," *IEEE Trans. Ind. Electron.*, vol. 65, no. 6, pp. 5023–5031, Jun. 2018.

[30] A. Anwar, W. Lin, X. Deng, J. Qiu, and H. Gao, "Quality inspection of remote radio units using depth-free image-based visual servo with acceleration command," *IEEE Trans. Ind. Electron.*, vol. 66, no. 10, pp. 8214–8223, Oct. 2019.

[31] C. Yang, P. Liu, G. Yin, H. Jiang, and X. Li, "Defect detection in magnetic tile images based on stationary wavelet transform," *NDT E Int.*, vol. 83, pp. 78–87, Oct. 2016.

[32] Y. Yang, Z.-J. Zha, M. Gao, and Z. He, "A robust vision inspection system for detecting surface defects of film capacitors," *Signal Process.*, vol. 124, pp. 54–62, Jul. 2016.

[33] Y. Li, D. Zhang, and D.-J. Lee, "Automatic fabric defect detection with a wide-and-compact network," *Neurocomputing*, vol. 329, pp. 329–338, Feb. 2019.

[34] S. Cheon, H. Lee, C. O. Kim, and S. H. Lee, "Convolutional neural network for wafer surface defect classification and the detection of unknown defect class," *IEEE Trans. Semicond. Manuf.*, vol. 32, no. 2, pp. 163–170, May 2019.

[35] M. Haselmann and D. P. Gruber, "Pixel-wise defect detection by CNNs without manually labeled training data," *Appl. Artif. Intell.*, vol. 33, no. 6, pp. 548–566, Mar. 2019.

[36] J. Li, H. He, L. Li, and G. Chen, "A novel generative model with bounded-GAN for reliability classification of gear safety," *IEEE Trans. Ind. Electron.*, vol. 66, no. 11, pp. 8772–8781, Nov. 2019.

[37] L. Itti, C. Koch, and E. Niebur, "A model of saliency-based visual attention for rapid scene analysis," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 20, no. 11, pp. 1254–1259, Nov. 1998.

[38] K. Xu *et al.*, "Show, attend and tell: Neural image caption generation with visual attention," in *Proc. 32nd Int. Conf. Mach. Learn.*, Jul. 2015, vol. 37, pp. 2048–2057. [Online]. Available: http://proceedings.mlr.press/v37/xuc15.html

[39] J. Hu, L. Shen, and G. Sun, "Squeeze-and-excitation networks," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 7132–7141.

[40] S. Woo, J. Park, J.-Y. Lee, and I. S. Kweon, "CBAM: Convolutional block attention module," in *Proc. Eur. Conf. Comput. Vis.*, Oct. 2018, pp. 3–19.

[41] U. Leonards, S. Sunaert, P. Van Hecke, and G. A. Orban, "Attention mechanisms in visual search—An fMRI study," *J. Cogn. Neurosci.*, vol. 12, no. Suppl. 2, pp. 61–75, Nov. 2000.

[42] F. Wang *et al.*, "Residual attention network for image classification," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jul. 2017, pp. 6450–6458.

[43] L. Guo, Y. Lei, S. Xing, T. Yan, and N. Li, "Deep convolutional transfer learning network: A new method for intelligent fault diagnosis of machines with unlabeled data," *IEEE Trans. Ind. Electron.*, vol. 66, no. 9, pp. 7316–7325, Sep. 2019.

[44] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," 2014, *arXiv:1409.1556*.

[45] X. Glorot, A. Bordes, and Y. Bengio, "Deep sparse rectifier neural networks," in *Proc. 4th Int. Conf. Artif. Intell. Statist.*, Apr. 2011, vol. 15, pp. 315–323. [Online]. Available: http://proceedings.mlr.press/v15/glorot11a.html

[46] J. Deng, W. Dong, R. Socher, L. Li, K. Li, and L. Fei-Fei, "ImageNet: A large-scale hierarchical image database," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2009, pp. 248–255.

**Chuanfei Hu** received the B.S. degree in electrical engineering and automation from the Jiangsu University of Science and Technology, Zhenjiang, China. He is currently working toward the M.S. degree in control engineering with the University of Shanghai for Science and Technology, Shanghai, China.

His research interests include computer vision and applications of deep learning.

**Yongxiong Wang** received the B.S. degree in engineering mechanics from Harbin Engineering University, Harbin, China, and the M.S. and Ph.D. degrees in control science and engineering from Shanghai JiaoTong University, Shanghai, China.

He is currently a Professor of Control Science and Engineering with the University of Shanghai for Science and Technology, Shanghai, China. His research interests include computer vision and intelligent robot.