



# Defect Detection Based on Improved YOLOx for Ultrasonic Images

Liangshan Lou<sup>1</sup> · Ke Lu<sup>1</sup> · Jian Xue<sup>1</sup>

Received: 25 February 2023 / Revised: 10 November 2023 / Accepted: 8 January 2024

© The Author(s), under exclusive licence to Springer Science+Business Media, LLC, part of Springer Nature 2024

## Abstract

Non-destructive testing is a method of detecting defects in materials or electronic components without causing damage to the detected objects. The most commonly used detection technology is ultrasonic detection. However, for images generated by ultrasonic inspection, manual recognition and traditional image processing methods are mostly used for defect identification, which are both inefficient and costly. The detection of defects in printed circuit boards (PCB) is a particularly difficult problem in the field of industrial inspection, which has strict requirements. We adopt a deep learning method to implement intelligent defect detection in our work. To address the lack of training data, we collect PCB surface images using a high-resolution ultrasonic microscope and create a dataset by annotating the defects in the images. The final dataset can be used for defect object detection based on deep learning. Furthermore, we propose an improved object detection method for defect detection that adopts the four-scale Swin Transformer as the multi-scale feature extraction network and uses the decoupled head from YOLOx to output defect categories and locations. To better learn the defect features, we pretrain on datasets of PCB images obtained using other methods, such as charge-coupled devices and CMOS sensors. Subsequently, we transfer to our own created dataset to perform training and testing. Experimental results show that our improved model achieves an average precision of 99.9% on our PCB test dataset, and an average precision of 85.1% on PASCAL VOC 2007 test dataset while extending to the conventional object detection.

**Keywords** Object detection · Defect detection · High-resolution ultrasonic microscope · Printed circuit board · Swin Transformer · YOLOx

---

Extended author information available on the last page of the article

Published online: 03 February 2024

Springer

## 1 Introduction

Non-destructive testing is a widely used technique in fields such as aerospace, industrial production, and chip processing. Because it does not damage the material, the non-destructive detection of possible defects improves the reliability of the production and operation of an enterprise and eliminates potential safety hazards. At present, commonly used nondestructive testing methods are laser, X-ray, and ultrasonic methods. However, among these methods, ultrasonic testing has gradually increased in use because it is safe, inexpensive, and has strong penetration. In contrast to other detection methods, ultrasound inspection obtains not only the surface image of the detected object, but also the internal image. Some of the advantages of ultrasonic testing include simple operation, precise defect localization [1] and the ability to evaluate the structure of the components according to different acoustic properties [2].

Ultrasound is emitted through an ultrasound probe, known as a transducer. The propagation speed of ultrasonic waves in different materials is different. Ultrasonic waves are very penetrative and can pass through the surface of the detected object. However, after they pass through the surface, reflections are generated because of the characteristics of the sound waves. According to the reflected signal, we can determine which layer the signal is coming from. In addition, different modes of data acquisition, A-scans, B-scans, or C-scans, are adopted. It is common to see a B-scan created from hundreds of A-scans. A C-scan is a three-dimensional imaging scan, and the scan result is a cross-section of the detected object. Other representations of ultrasonic testing data, such as volume-corrected B-scans [3] and D-scans [3], are also frequently used.

In recent years, ultrasound imaging technology has matured, the resolution of images has increased, and the method of acquiring ultrasonic data has become simpler. However, there are very few studies on the further processing of the acquired data. For example, when the image of an object to be inspected is obtained by an ultrasonic imaging device, manual methods based on the operator's experience are used, or traditional image processing methods are used to roughly identify defects in the detected objects. However, manual detection can lead to false and missed detections, and it requires a large amount of human and financial resources. There are still very few studies on automated defect localization and identification based on ultrasonic images, which limits the application of ultrasonic defect detection. The advent of machine learning has led researchers to consider how to implement automatic defect detection. The most commonly used method in the field of non-destructive testing is based on the analysis of waveform data from reflected ultrasonic waves. The coefficients obtained using wavelet, Fourier, or cosine transforms are used as the input to traditional classifiers, such as artificial neural networks [4], support vector machine and decision tree, which then determine whether the detected object has defects. The data used in these methods are mostly obtained by A-scans. Such data only reflect whether the signal at a certain position is abnormal, and the information of the surrounding context cannot be referenced, so the detected result is relatively unreliable. To

better determine whether the detected object has defects, researchers have begun to use images obtained by B-scans. Images from a B-scan contain more spatial information than the waveform data from an A-scan.

An artificial neural network framework was proposed by Yuan et al. [5] to identify the echoes from steel train wheels by B-scans. The whole network is divided into two parts: the first part identifies whether the signal is from noise, and the second part determines whether the echoes are from defects or not. To solve the problem of insufficient ultrasonic data for training, Virkkunen et al. [3] collected ultrasonic data using B-scans, used data augmentation to enhance the limited raw data, and then adopted a deep convolutional neural network (CNN) to detect flaws from phased-array ultrasonic data, which proved that deep learning methods could detect defects. Posilovi et al. [6] collected ultrasonic images by B-scan from steels, and used YOLOv1 [7] and single shot multibox detector (SSD) [8] independently to detect defects in images, which obtained high mean average precision. Their results proved that typical object detection algorithms could be used to detect defects. As a result, Medak et al. [9] further compared the performance of different object detection algorithms for defect detection in ultrasonic images, and the EfficientDet [10] series of methods were found to outperform the other deep learning models YOLOv3 [11] and RetinaNet [12]. Meng et al. [13] developed a CNN network for the automation of ultrasonic signal classification from C-scan signals in a carbon fiber reinforced polymer structure. However, almost no researchers have implemented defect detection based on deep learning for ultrasonic C-scan images.

In our work, we choose printed circuit boards (PCBs) as the detection target, mainly because PCB defect detection is in high demand in the industry. Since the images obtained by charge coupled devices (CCDs) are easily affected by problems such as light and angle changes, image acquisition using a high-resolution ultrasonic microscope can completely solve these issues. However, there is currently no research on PCB defect detection based on deep learning for ultrasonic images. To address the lack of ultrasonic training data for defect detection based on deep learning, we use PCB test samples to collect surface images by C-scans, annotate them and construct an ultrasonic dataset. We also use the current state-of-the-art object detection model—YOLOx [14]—which outperformed the YOLO series in 2021, and further improve the defect detection capability of the model by replacing the feature extraction network with the four-scale Swin Transformer [15]. Experimental results show that our improved defect detection model achieves an average precision of 99.9% on our PCB test dataset, much higher than the compared models, including Faster R-CNN [16], EfficientDet [10], YOLOv3 [11] and YOLOv4 [17]. While extending to the conventional object detection, our model also achieves better detection results on PASCAL VOC 2007 test dataset, compared with the existing state-of-the-art models. In summary, our contributions can be summed up as follows:

- We use a high-resolution ultrasonic microscope, which is also the equipment independently developed by our laboratory, to inspect the PCB surface by C-scan, generate a batch of raw ultrasonic images, and annotate them. To compensate for the lack of training data for deep learning, we collect a total of 200 original images and expand them to 4320 images through data augmentation.

They are divided into training and test datasets to provide both a public dataset and a benchmark for future research in the field of PCB defect detection.

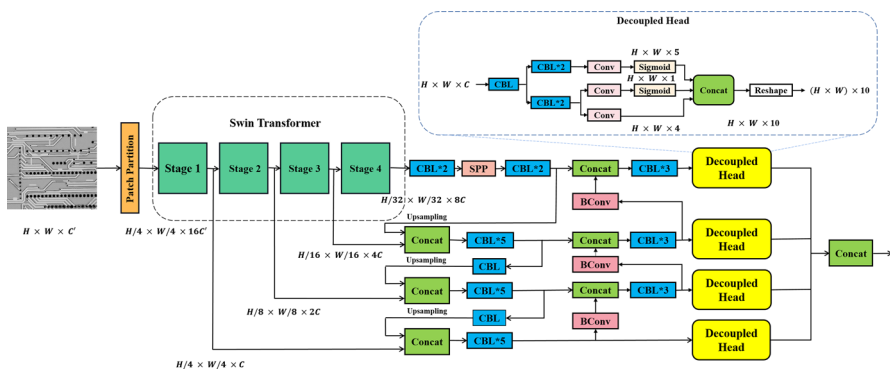
- We replace the original feature extraction network in YOLOx with a Swin Transformer pretrained on ImageNet to enhance the feature expression ability and multi-scale feature fusion ability of the network. Instead of the three scales of YOLOx, we adopt four-scale Swin Transformer while retaining the YOLOx decoupled head to output the prediction results of defect categories and locations, enhancing the capability of detecting small defect objects.
- To further address the lack of data, we also pretrain the model on the public PCB datasets collected using other methods, such as CCDs and CMOS sensors, to prevent over-fitting and improve generalization. We then continue training the above pretrained model on our ultrasonic PCB dataset.

## 2 Methodology

The architecture of the proposed defect detection model is shown in Fig. 1. As can be seen in the figure, our model is an end-to-end model that adopts Swin Transformer as the feature extraction network. The decoupled heads from YOLOx output the prediction results corresponding to different scales, and we concatenate them as the final prediction results.

### 2.1 A Brief View of Object Detection Methods

At present, there are many conventional object detection methods based on deep learning, and they are divided into two main categories: one-stage and two-stage object detection methods. In one-stage object detection methods, it is not necessary



**Fig. 1** Architecture of our model with four-scale Swin Transformer. We fuse multi-scale features from four scales, instead of three scales. A CBL block from YOLOx is composed of a convolutional layer, batch normalization layer and leaky-relu activation function. A SPP block is composed of a CBL block and multiple maximum pooling, the outputs of which are concatenated at the end. A BConv represents a convolutional layer to implement downsampling. The fused features from four scales are input into the decoupled heads, respectively

to obtain a proposed bounding box. Usually, the image is divided into dense grids and each grid is responsible for determining whether it includes an object. Then, the category probability and position coordinates of the object are directly generated. In this way, the final detection result is obtained after a single detection. The speed of such object detection methods is generally faster than that of the two-stage algorithms, but the accuracy is relatively lower. Typical one-stage detector algorithms include YOLOv1, YOLOv2 [18], YOLOv3, YOLOv4, and YOLOx. In addition to YOLO series models, other object detection algorithms have been proposed, such as SSD and EfficientDet [10] which is a series models that balance accuracy and speed. In two-stage object detection methods, the first stage focuses on determining where the object appears, obtaining the proposed bounding box using a region proposal network (RPN) [16], and ensuring sufficient accuracy and recall, and the second stage focuses on classifying the object in the proposal box and determining more precise locations. Such object detection methods are usually more accurate, but slower. Typical two-stage detector algorithms include R-CNN [19], SPP-Net [20], Fast R-CNN [21] and Faster R-CNN [16].

## 2.2 Main Architecture of the Network

We consider existing object detection algorithms and select YOLOx to implement defect detection. YOLOx has achieved state-of-the-art performance with respect to both inference speed and prediction accuracy. That is why we choose YOLOx as our backbone. However, CSPDarkNet [14] used in YOLO series is a multi-layer CNN, which can not take contextual information into account well, and contextual information is crucial for determining the properties of an object in an image. Under the premise of only considering the accuracy, we use Swin Transformer to replace CSPDarkNet as the feature extraction network. Swin Transformer is based on ViT [22] by introducing priors such as hierarchy, locality and translation invariance, and achieves better performance in visual tasks. In addition, its design, which incorporates a shifted window, means its complexity is linearly related to the image size, and its computational efficiency is very good. Swin Transformer consists of four similar stages, each of which has different numbers of Swin Transformer blocks. A Swin Transformer block [15] consists of a shifted window-based multi-head self-attention module, followed by a two-layer multilayer perception with a GELU activation function between them.

First, we use an image of size  $H \times W \times C'$  as the input, where  $H$ ,  $W$  and  $C'$  represent height, width and channels, and then divide it into equal-sized patches by patch partition. We can get the original features with a size of  $\frac{H}{4} \times \frac{W}{4} \times 16C'$ . After patch embedding, the size of the feature map becomes  $\frac{H}{4} \times \frac{W}{4} \times C$ . Each time a stage after the initial stage is passed, the height  $H$  and width  $W$  are halved, and the number of channels  $C$  is doubled. After processed by four stages, we can get four feature maps, the sizes of which are  $\frac{H}{4} \times \frac{W}{4} \times C$ ,  $\frac{H}{8} \times \frac{W}{8} \times 2C$ ,  $\frac{H}{16} \times \frac{W}{16} \times 4C$  and  $\frac{H}{32} \times \frac{W}{32} \times 8C$ , respectively. Instead of the three scales used by YOLOx, we use four scales by additionally adding the output of the first high-resolution stage. We do this mainly

because the size of a defect is usually small, and one more high-resolution scale helps us better obtain defect information.

First, we adopt the same design to fuse different-scale features from top to down to build a feature pyramid network (FPN). Here, we use  $X_i$  to represent the output of the  $i$ -th stage, and  $D_i$  to represent the input of the  $i$ -th decoupled head, which is defined as follows [14]:

$$\begin{aligned} D_i &= 5 * \text{CBL}(\text{Concat}(X_i, \text{Upsampling}(\text{CBL}(D_{i+1})))) \\ D_4 &= 2 * \text{CBL}(\text{SPP}(2 * \text{CBL}(X_4))) \end{aligned} \quad (1)$$

where a CBL block [14] from YOLOx represents a convolutional layer, batch normalization layer and leaky-relu activation function, and a SPP block [14] from YOLOx is composed of a CBL block and multi max-pooling, the outputs of which are concatenated at the end. After that, we implement the feature fusion from bottom to up to build a path-aggregation network (PANet). We use  $P_i$  to represent the input of the  $i$ -th decoupled head, which is defined as follows:

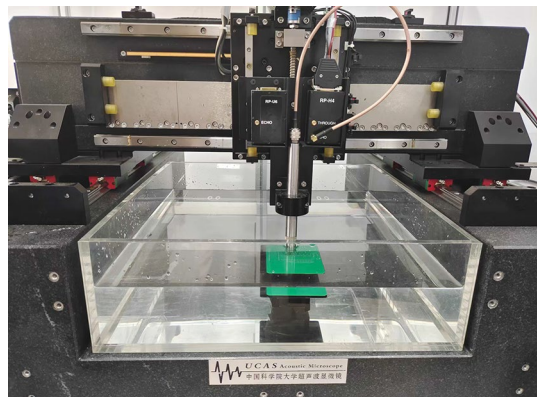
$$P_i = 3 * \text{CBL}(\text{Concat}(D_i, \text{BConv}(P_{i-1}))) \quad (2)$$

where BConv represents a convolutional layer to implement downsampling. The fused feature of each scale is then processed by a decoupled head [14], as shown in Fig. 1, which can be decoupled into three types of information: classes, objects, and boxes. At last, the outputs of four decoupled heads are concatenated as the final output.

## 2.3 Training

To address the problem of insufficient ultrasonic image data, we use a high-resolution ultrasonic microscope, as shown in Fig. 2, and the C-scan method to scan the surface of the two-layer PCB with no solder joints. Most PCB traces are less than 1 mm in width. Currently, available ultrasonic frequencies in our laboratory are 30

**Fig. 2** High-resolution ultrasonic microscope, which is the equipment independently developed by our laboratory



MHZ, 50 MHZ, and 180 MHZ. In our experiment, a 30 MHZ probe is used, the constant temperature of distilled water is 25 °C, the speed of sound in water is 1480 m/s, the speed of sound in copper is 4700 m/s, and the time resolution is 0.001  $\mu$ s. We use 20 different PCBs as samples, and scan five locations on each board: the four corners and the center region, resulting in five different images of  $800 \times 800$  pixels. Finally, we obtain 200 original ultrasonic images, the details of which are listed in Table 1. Because there are few defects in the natural state, we artificially add some common defects according to the practice in [23]. The PCB defects include five classes (mouse bite, open circuit, short, spur and spurious copper), as shown in Fig. 3. We randomly crop our images into  $640 \times 640$ . Then, we use data augmentation to expand our dataset, including flip, rotation, blur and noise. Additional details are listed in Table 2. According to the ratio of 8:2, we randomly split our dataset into training and test dataset, obtaining 3456 images for training and 864 images for testing.

Since Transformer has less inductive bias than CNN, it has poor performance when the dataset is small. To enhance the feature extraction capability of our improved model and its generalization, it is necessary to introduce pretraining and transfer learning. Our model with the pre-trained Swin Transformer on ImageNet-22k dataset [24] that we use as the feature extraction network firstly is pre-trained on MSCOCO 2017 dataset [25]. Although there exists almost no research in the field of defect detection on PCB ultrasonic images, a few studies related to

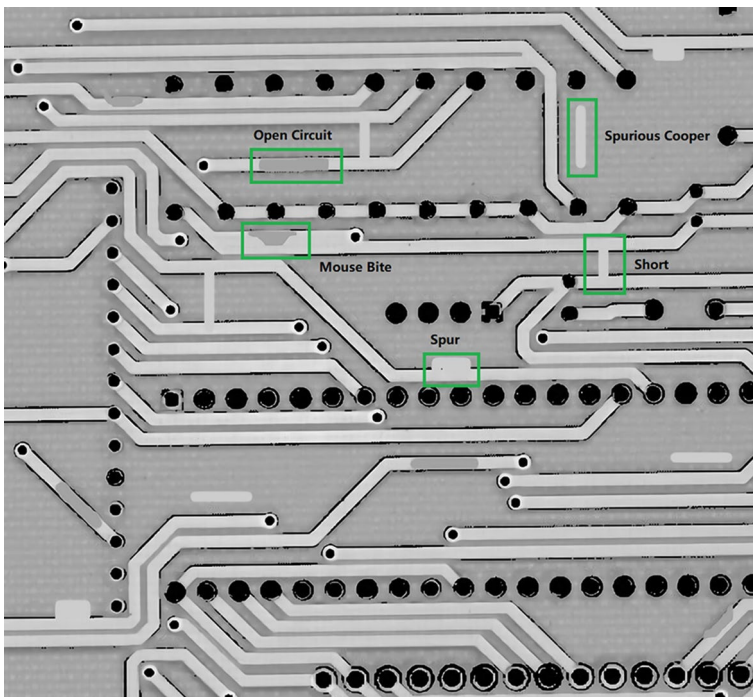
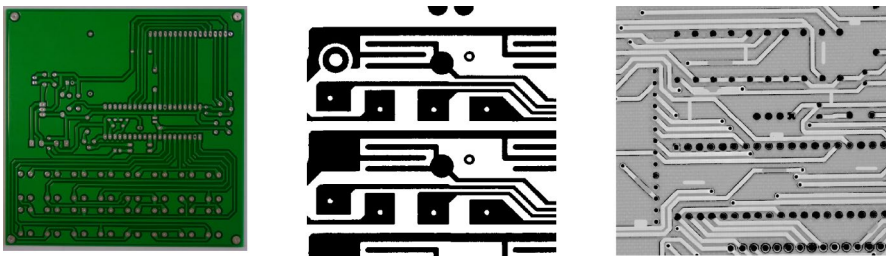


Fig. 3 Example of defects

PCBs are based on images collected by CCDs or CMOS, such as the public dataset created by Huang et al. [26], and another public dataset created by Tang et al. [27]. Because the defects in these PCB datasets are almost the same as those in ours, we continue to pre-train the above obtained model on these CCD-based and CMOS-based PCB datasets and then transfer to our ultrasonic dataset for training. Figure 4 compares example images from our ultrasonic dataset with those from the other two datasets. As can be seen in the figure, these images are very similar in nature, and that is why pretraining is performed on CCD-based and CMOS-based PCB datasets. Furthermore, to evaluate the performance of our improved model on the conventional object detection, we also train our model on PASCAL VOC 2007 and 2012 dataset after pretrained on MSCOCO 2017 dataset, and then test its performance on PASCAL VOC 2007 test dataset.



**Fig. 4** Comparison of our ultrasonic dataset with the other two datasets scanned from CCDs and CMOS sensors. The left CMOS-based example is from Huang et al. [26], the middle CCD-based example is from Tang et al. [27] and the right is ours

**Table 1** Original dataset overview

Table	Number of detects	Number of images
Mouse bite	105	78
Open circuit	115	81
Short	129	97
Spur	228	129
Spurious copper	104	85
Total	681	200

**Table 2** Augmented dataset overview

Table	Number of detects	Number of images
Mouse bite	2214	1800
Open circuit	2430	1890
Short	2712	2280
Spur	4788	3030
Spurious copper	2202	1986
Total	14,346	4320



For data preprocessing, we adopt the data augmentation used in YOLOx, including Mosaic [11] and Mixup [28], which have been proven to efficiently improve performance, especially the performance of small object detection. In addition, we continue to use SimOTA [14, 29] used in YOLOx for positive sample screening. For the loss function, a multi-task loss composed of three losses is used, which include coordinate regression loss, object loss, and classification loss. The object loss is responsible for determining whether the target box is an object, and based on that information, the classification loss is responsible for determining which class the object belongs to. The above two losses both adopt binary cross-entropy (BCE) loss function, which is defined as follows:

$$Loss_{BCE} = -\omega[y \log p_\theta + (1 - y) \log(1 - p_\theta)] \quad (3)$$

where  $\omega$  is a hyperparameter,  $y$  is the ground truth,  $\theta$  is the model parameter, and  $p_\theta$  is the output of the model. Usually, the above activation function adopts sigmoid. For the coordinate regression loss, the intersection over union (IoU) is adopted, which is defined as follows:

$$Loss_{IoU} = -\log \frac{BB \cap GT}{BB \cup GT} \quad (4)$$

where  $GT$  represents ground truth and  $BB$  represents the predicted results.

### 3 Experiments and Results

#### 3.1 Dataset

In order to implement defect detection, we adopt our own collected dataset to train our model, where 3456 images are used for training and 864 images for testing. Before that, we conduct pretraining twice. First, we pretrain our model on MSCOCO 2017, which includes 118,287 images for training, 5000 images for validation and 40,670 images for testing. Then, we continue to pretrain the above model on the following two PCB datasets from CCDs and CMOS: the one dataset made by Huang et al. [26], which includes 8534 images for training and 2134 images for testing, and the other dataset made by Tang et al. [27], which includes 1000 images for training and 500 images for testing.

To further prove the performance of our improved model on the conventional object detection, we train and test our model on PASCAL VOC 2007 dataset, which includes 5011 images for training and validation and 4952 images for testing, and PASCAL VOC 2012 dataset, which includes 11540 images for training and validation and 10,991 images for testing.

#### 3.2 Experimental Settings

The mean average precision metrics is used for the performance evaluation, as given in the later versions of PASCAL VOC (2010-2012) [30], which is the commonly

used evaluation metrics for object detection. Here,  $\mathbf{AP}_{50}$  indicates average precision when the value of the IoU is set to 0.5.  $\mathbf{AP}$  indicates the average precision using IoU values ranging from 0.5 to 0.95 with a step size of 0.05.

We set the number of stages in Swin Transformer to 4, and patch size to 4. The window size is set to 20, and each window of the high-resolution feature map is  $80 \times 80$  pixels on the original image, which is enough to cover small defects. The pretrained model of Swin Transformer we select is pretrained on the image size of  $384 \times 384$ . Since our image size is  $640 \times 640$  which mismatches the pretrained image size, we use geometric interpolation to solve this problem. In addition, since the number of images in our dataset is limited, we set a maximum of 300 epochs for training to prevent over-fitting. The augmentation of Mosaic and MixUp remains open until the last 15 epochs [14]. We set the batch size to 8, and adopt half-precision training. We adopt stochastic gradient descent as the optimizer with an original learning rate of 0.00125, and the cosine learning mechanism is adopted. We set the momentum parameter to 0.9 and weight decay to  $5 \times 10^{-4}$ .

### 3.3 Ablation Study

To evaluate the effect of our model improvements, we conduct ablation experiments. First, we compare the difference between YOLOx and our improved model before and after pretraining. The results of ablation study are presented in Table 3. When our improved model uses Swin Transformer with three scales, the prediction accuracy after pretraining is **99.8%  $\mathbf{AP}_{50}$**  and **82.0%  $\mathbf{AP}$** , which is **1.7%  $\mathbf{AP}_{50}$**  and **2.0%  $\mathbf{AP}$**  higher than that before pretraining. It proves that when there are insufficient data, pretraining can improve the model performance to a certain extent. As shown in the table, the prediction accuracy of our improved model with four-scale Swin Transformer achieves **99.9%  $\mathbf{AP}_{50}$**  and **83.1%  $\mathbf{AP}$** , higher than models with CSPDarkNet and three-scale Swin Transformer, which proves that one more high-resolution scale can help to improve the accuracy of small defect detection. Furthermore, we continue to conduct ablation study on PASCAL VOC 2007 test dataset, as shown in Table 4, and the comparison results further prove the effect of our model improvements. These results also prove that integrating the Swin Transformer into the YOLOx model improves the model's performance, although this requires a large number of data for training and the inference speed is slightly slower than that of the original YOLOx. In some cases where the accuracy required for defect detection

**Table 3** Ablation study of different backbones, scales and training methods on our ultrasonic dataset

Pretraining	Backbone	Scales	$\mathbf{AP}_{50}$ (%)	$\mathbf{AP}$ (%)
No	CSPDarkNet	3	98.8	80.8
Yes	CSPDarkNet	3	99.3	81.6
Yes	CSPDarkNet	4	99.4	81.8
No	Swin Transformer	3	98.1	80.0
Yes	Swin Transformer	3	99.8	82.0
Yes	Swin Transformer	4	<b>99.9</b>	<b>83.1</b>

The bold indicates the best performance

**Table 4** Ablation study of different backbones and scales on PASCAL VOC 2007 test dataset

Backbone	Scales	AP <sub>50</sub> (%)	AP (%)
CSPDarkNet	3	83.9	64.1
Swin Transformer	3	84.8	65.5
Swin Transformer	4	<b>85.1</b>	<b>65.9</b>

The bold indicates the best performance

is high, a YOLOx model using the Swin Transformer is a suitable solution. When high speed defect detection is required, we can use the original YOLOx, which also provides a good level of prediction accuracy. All the following experiments adopt four-scale Swin Transformer.

### 3.4 Performance Comparison

First, to evaluate the performance of our proposed method using the Swin Transformer, we train and test some typical object detection algorithms on our dataset, including Faster R-CNN [16], EfficientDet [10], YOLOv3 [11] and YOLOv4 [17], which all have once been proved to be efficient in the field of defect detection. Then, we compare our proposed model with the above models. For fair comparison, all the compared models are trained for 300 epochs. The results of performance comparison are presented in Table 5. As these results reveal, YOLOx with CSPDarkNet surpasses the other models on the evaluation metrics of both AP<sub>50</sub> and AP. We note that the inference speed of YOLOx with CSPDarkNet, which has been evaluated, is faster than the inference speed of the other compared models, and the average inference time of our method is 17ms per image, which is slightly slower than 14.7ms of YOLOx and viable for on-line testing. However, without considering inference speed, our improved model with Swin Transformer achieves the state-of-the-art performance, obtaining **99.9% AP<sub>50</sub>** and **83.1% AP**, which is **0.6% AP<sub>50</sub>** and **1.5% AP** higher, respectively, than the performance of YOLOx with CSPDarkNet. Figure 5 presents some visual examples of defect detection in the PCB ultrasonic images obtained by our model with the Swin Transformer. As is shown in the figure, our proposed model can precisely locate the defect and determine its category.

**Table 5** Performance comparison on our ultrasonic test dataset

Model	Backbone	Size	AP <sub>50</sub> (%)	AP (%)
Faster R-CNN	ResNet 101 [31]	640	94.5	57.9
EfficientDet-D0	EfficientNet-B0 [32]	640	97.1	72.3
EfficientDet-D7	EfficientNet-B7 [32]	640	98.5	80.1
YOLOv3	DarkNet-53 [11]	416	97.3	78.8
YOLOv4	CSPDarkNet-53	416	98.2	79.6
YOLOx	CSPDarkNet-53	640	99.3	81.6
Ours	Swin Transformer	640	<b>99.9</b>	<b>83.1</b>

The bold indicates the best performance

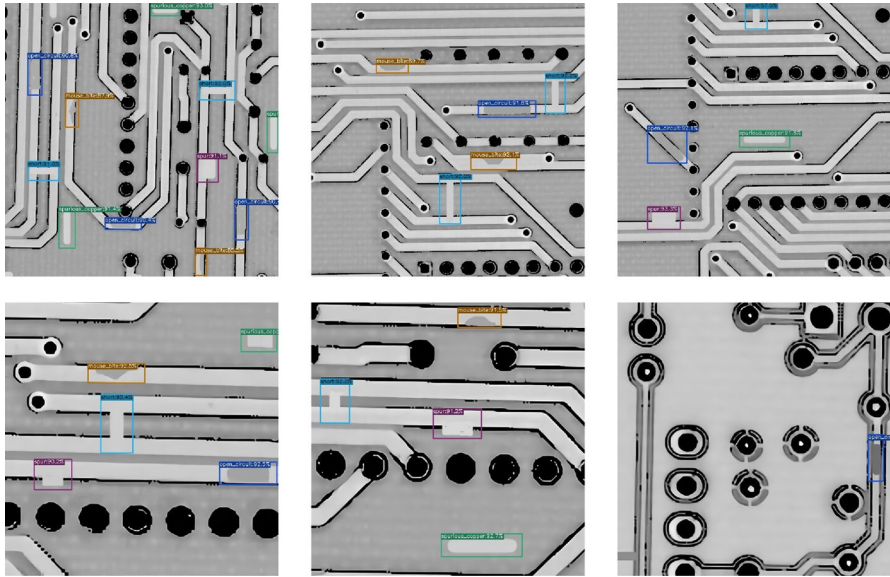


Fig. 5 Examples of detected defects on our dataset by our model

To further evaluate the conventional object detection performance of our proposed method, we train our model on PASCAL VOC 2007 and 2012 dataset [30], and then compare the performance of our improved model with that of the existing state-of-the-art models: HSD [33], Localize [34], EEEA-Net [35], and MViT [36]. All the compared models are tested on the PASCAL VOC 2007 test dataset. The comparison results are listed in Table 6. Because the integration of the Transformer into the model increases its complexity, reducing the corresponding inference speed, we only compare the prediction accuracy. As can be seen in the table, our improved model with four-scale Swin Transformer, pre-trained on MSCOCO, achieves the highest scores of **85.1% AP<sub>50</sub>** and **65.9% AP** on PASCAL VOC 2007 test dataset, which is **1.2% AP<sub>50</sub>** and **1.8% AP** higher, respectively, than the performance of YOLOx with CSPDarkNet. In addition, although MViT also achieves high average precision values, it is also composed of multi-modal ViT, which has lower inference

**Table 6** Performance comparison on PASCAL VOC 2007 test dataset with SOTA models from the leaderboard

Model	Backbone	Size	AP <sub>50</sub> (%)	AP (%)
HSD	VGG16 [37]	320	81.7	–
HSD	VGG16 [37]	512	83.0	–
Localize	RFBNet [38]	320	81.5	–
EEEA-Net	YOLOv4	320	81.8	–
MViT	ViT [22]	512	84.2	64.5
YOLOx	CSPDarkNet-53	640	83.9	64.1
Ours	Swin Transformer	640	<b>85.1</b>	<b>65.9</b>

The bold indicates the best performance



**Fig. 6** Examples of detected objects on PASCAL VOC 2007 by our model

speed than YOLOx with CSPDarkNet. Moreover, our improved model with four-scale Swin Transformer outperforms MViT with respect to both  $AP_{50}$  and  $AP$ . Figure 6 presents some visual examples of objects detected by our improved model on PASCAL VOC 2007 test dataset.

## 4 Conclusion

Defect detection based on ultrasonic images is still dominated by manual detection, which is a time-consuming and labor-intensive process, and prone to human error. Implementing the automated detection of defects is a very meaningful work. In our work, to address the serious shortage of PCB ultrasonic image data, we scan PCB using a high-resolution ultrasonic microscope, and create a PCB dataset based on ultrasonic images for deep learning model training and testing to implement the automated defect detection. These data can support and act as a benchmark for subsequent research in the field of PCB defect detection. Moreover, we demonstrate that pretraining on similar datasets can address the problem of insufficient training data to a certain extent and improve the performance of the model. We find that the YOLOx model can efficiently and accurately detect defects in ultrasonic images, and the accuracy of PCB defect detection has reached **99.3%  $AP_{50}$** . This also proves the feasibility of deep learning methods in defect detection for ultrasonic images. Without considering the complexity of the model, we use four-scale Swin Transformer instead of CSPDarkNet in YOLOx, and the accuracy of PCB defect detection has reached **99.9%  $AP_{50}$**  and **83.1%  $AP$** . In the next step of our research, we will continue to collect data to expand our dataset, and then use our equipment and algorithms to implement internal defect detection.

**Author Contributions** LL and JX proposed the idea. LL, KL, and JX completed the methodology design and model creation and performed the validation. LL conducted the investigation and formal analysis. KL and JX provided the experimental resources. LL completed data collection and annotation and wrote the original draft. JX was responsible for review and editing. This work was under the supervision of KL and JX. All authors have read and reviewed the manuscript and agreed to the published version of the manuscript.

**Funding** This work was supported by the National Natural Science Foundation of China (61929104, 61972375), R & D Program of Beijing Municipal Education Commission (2019022), Scientific Research Program of Beijing Municipal Education Commission (KZ201911417048) and Beijing Natural Science Foundation (4212011).

**Data Availability** The PASCAL VOC 2007 and 2012 dataset [30] is publicly available at <http://host.robots.ox.ac.uk/pascal/VOC/>, the MSCOCO 2017 dataset [25] at [www.cocodataset.org/](http://www.cocodataset.org/), CMOS-based PCB dataset [26] at <https://robotics.pkusz.edu.cn/resources/dataset/>, and CCD-based PCB dataset [27] at <https://github.com/tangsanli5201/DeepPCB>. Our ultrasonic dataset has been provided publicly at <https://iiplab.net/u-pcbd/>.

## Declarations

**Ethical approval** Not applicable.

**Conflict of interest** The authors have no competing interests to declare that are relevant to the content of this article.

## References

1. Ye, J., Ito, S., & Toyama, N. (2018). Computerized ultrasonic imaging inspection: From shallow to deep learning. *Sensors (Basel, Switzerland)*, 18(11), 3820.
2. Davi, S., Mineo, C., Macleod, C., Pierce, S. G., & Mccubbin, C. (2020). Correction of b-scan distortion for optimum ultrasonic imaging of backwalls with complex geometries. *Insight-Non-Destructive Testing and Condition Monitoring*, 62(4), 184–191.
3. Virkkunen, I., Koskinen, T., Jessen-Juhler, O., & Rinta-Aho, J. (2019). Augmented ultrasonic data for machine learning. *abs/1903.11399*.
4. Bettayeb, F., Rachedi, T., & Benbartaoui, H. (2004). An improved automated ultrasonic NDE system by wavelet and neuron networks. *Ultrasonics*, 42(1), 853–858.
5. Yuan, M., Li, J., Liu, Y., & Gao, X. (2020). Automatic recognition and positioning of wheel defects in ultrasonic b-scan image using artificial neural network and image processing. *Journal of Testing and Evaluation*, 48(1), 20180545.
6. Posilovic, L., Medak, D., Subasic, M., Petkovic, T., Budimir, M., & Loncaric, S. (2019). Flaw detection from ultrasonic images using YOLO and SSD. In *11th international symposium on image and signal processing and analysis* (pp. 163–168). <https://doi.org/10.1109/ISPA.2019.8868929>.
7. Redmon, J., Divvala, S. K., Girshick, R. B., & Farhadi, A. (2016). You only look once: Unified, real-time object detection. In *2016 IEEE conference on computer vision and pattern recognition* (pp. 779–788). <https://doi.org/10.1109/CVPR.2016.91>.
8. Liu, W., Anguelov, D., Erhan, D., Szegedy, C., Reed, S. E., Fu, C., & Berg, A. C. (2016) SSD: Single shot multibox detector. In *Computer vision—ECCV 2016—14th European conference* (Vol. 9905, pp. 21–37).
9. Medak, D., Posilović, L., Subašić, M., Budimir, M., & Lončarić, S. (2021). Automated defect detection from ultrasonic images using deep learning. *IEEE Transactions on Ultrasonics, Ferroelectrics, and Frequency Control*, 68(10), 3126–3134. <https://doi.org/10.1109/TUFFC.2021.3081750>
10. Tan, M., Pang, R., & Le, Q. V. (2020). Efficientdet: Scalable and efficient object detection. In *2020 IEEE/CVF conference on computer vision and pattern recognition* (pp. 10778–10787). <https://doi.org/10.1109/CVPR42600.2020.01079>.
11. Redmon, J., & Farhadi, A. (2018). Yolov3: An incremental improvement. *CoRR* **abs/1804.02767**.
12. Lin, T., Goyal, P., Girshick, R. B., He, K., & Dollár, P. (2017) Focal loss for dense object detection. In *IEEE international conference on computer vision* (pp. 2999–3007). <https://doi.org/10.1109/ICCV.2017.324>
13. Meng, M., Chua, Y. J., Wouterson, E., & Ong, C. P. K. (2017). Ultrasonic signal classification and imaging system for composite materials via deep convolutional neural networks. *Neurocomputing*, 257, 128–135.
14. Ge, Z., Liu, S., Wang, F., Li, Z., & Sun, J. (2021). YOLOX: Exceeding YOLO series in 2021. *CoRR* **abs/2107.08430**.

15. Liu, Z., Lin, Y., Cao, Y., Hu, H., Wei, Y., Zhang, Z., Lin, S., & Guo, B. (2021). Swin transformer: Hierarchical vision transformer using shifted windows. In *2021 IEEE/CVF international conference on computer vision* (pp. 9992–10002). <https://doi.org/10.1109/ICCV48922.2021.00986>.
16. Ren, S., He, K., Girshick, R. B., & Sun, J. (2015). Faster R-CNN: Towards real-time object detection with region proposal networks. In *Advances in neural information processing systems 28: Annual conference on neural information processing systems 2015* (pp. 91–99).
17. Bochkovskiy, A., Wang, C., & Liao, H. M. (2020). Yolov4: Optimal speed and accuracy of object detection. *CoRR abs/2004.10934*.
18. Redmon, J., & Farhadi, A. (2017). YOLO9000: Better, faster, stronger. In *2017 IEEE conference on computer vision and pattern recognition* (pp. 6517–6525). <https://doi.org/10.1109/CVPR.2017.690>.
19. Girshick, R. B., Donahue, J., Darrell, T., & Malik, J. (2014). Rich feature hierarchies for accurate object detection and semantic segmentation. In *2014 IEEE conference on computer vision and pattern recognition* (pp. 580–587). <https://doi.org/10.1109/CVPR.2014.81>.
20. He, K., Zhang, X., Ren, S., & Sun, J. (2014). Spatial pyramid pooling in deep convolutional networks for visual recognition. In *Computer vision—ECCV 2014—13th European conference* (Vol. 8691, pp. 346–361). [https://doi.org/10.1007/978-3-319-10578-9\\_23](https://doi.org/10.1007/978-3-319-10578-9_23).
21. Girshick, R. B. (2015). Fast R-CNN. In *2015 IEEE international conference on computer vision, ICCV 2015* (pp. 1440–1448). <https://doi.org/10.1109/ICCV.2015.169>.
22. Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J., & Houlsby, N. (2020) An image is worth 16x16 words: Transformers for image recognition at scale. *CoRR abs/2010.11929*.
23. Ding, R., Dai, L., Li, G., & Liu, H. (2019). TDD-net: A tiny defect detection network for printed circuit boards. *CAAI Transactions on Intelligence Technology*, 4(2), 110–116.
24. Krizhevsky, A., Sutskever, I., & Hinton, G. E. (2017). Imagenet classification with deep convolutional neural networks. *Communications of the ACM*, 60(6), 84–90.
25. Lin, T., Maire, M., Belongie, S. J., Hays, J., Perona, P., Ramanan, D., Dollár, P., Zitnick, C. L. (2014). Microsoft COCO: Common objects in context. In *Computer vision—ECCV 2014—13th European conference* (Vol. 8693, pp. 740–755).
26. Huang, W., & Wei, P. (2019). A PCB dataset for defects detection and classification. *CoRR abs/1901.08204*.
27. Tang, S., He, F., Huang, X., & Yang, J. (2019). Online PCB defect detector on A new PCB defect dataset. *CoRR abs/1902.06197*.
28. Zhang, H., Cissé, M., Dauphin, Y. N., & Lopez-Paz, D. (2018). mixup: Beyond empirical risk minimization. In *6th international conference on learning representations*.
29. Ge, Z., Liu, S., Li, Z., Yoshie, O., & Sun, J. (2021). OTA: optimal transport assignment for object detection. In *IEEE conference on computer vision and pattern recognition* (pp. 303–312).
30. Everingham, M., Eslami, S. M. A., Gool, L. V., Williams, C. K. I., Winn, J. M., & Zisserman, A. (2015). The pascal visual object classes challenge: A retrospective. *International Journal of Computer Vision*, 111(1), 98–136.
31. He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep residual learning for image recognition. In *2016 IEEE conference on computer vision and pattern recognition* (pp. 770–778).
32. Tan, M., & Le, Q. V. (2019). Efficientnet: Rethinking model scaling for convolutional neural networks. In *Proceedings of the 36th international conference on machine learning* (Vol. 97, pp. 6105–6114).
33. Cao, J., Pang, Y., Han, J., & Li, X. (2019). Hierarchical shot detector. In *2019 IEEE/CVF international conference on computer vision* (pp. 9704–9713).
34. Zhang, H., Fromont, É., Lefèvre, S., & Avignon, B. (2020). Localize to classify and classify to localize: Mutual guidance in object detection. In *Computer vision—ACCV 2020—15th Asian conference on computer vision* (Vol. 12625, pp. 104–118).
35. Termritthikun, C., Jamtsho, Y., Ieamsaard, J., Muneesawang, P., & Lee, I. (2021). EEEA-Net: An early exit evolutionary neural architecture search. *Engineering Applications of Artificial Intelligence*, 104, 104397.
36. Maaz, M., Rasheed, H. A., Khan, S. H., Khan, F. S., Anwer, R. M., & Yang, M. (2021). Multi-modal transformers excel at class-agnostic object detection. *CoRR abs/2111.11430*.
37. Simonyan, K., & Zisserman, A. (2015). Very deep convolutional networks for large-scale image recognition. In *3rd international conference on learning representations*.
38. Liu, S., Huang, D., & Wang, Y. (2018). Receptive field block net for accurate and fast object detection. In *Computer vision—ECCV 2018—15th European conference* (Vol. 11215, pp. 404–419).

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

Springer Nature or its licensor (e.g. a society or other partner) holds exclusive rights to this article under a publishing agreement with the author(s) or other rightsholder(s); author self-archiving of the accepted manuscript version of this article is solely governed by the terms of such publishing agreement and applicable law.

## Authors and Affiliations

Liangshan Lou<sup>1</sup> · Ke Lu<sup>1</sup> · Jian Xue<sup>1</sup>

✉ Jian Xue  
xuejian@ucas.ac.cn

Liangshan Lou  
louliangshan20@mails.ucas.ac.cn

Ke Lu  
luk@ucas.ac.cn

<sup>1</sup> School of Engineering Science, University of Chinese Academy of Sciences, Beijing 100049, China