International Journal of Computational Intelligence Systems (2024) 17:29
https://doi.org/10.1007/s44196-024-00423-w

RESEARCH ARTICLE

Deep Learning-Based Integrated Circuit Surface Defect Detection: Addressing Information Density Imbalance for Industrial Application

Xiaobin Wang[1] · Shuang Gao[1] · Jianlan Guo[1] · Chu Wang[2] · Liping Xiong[1] · Yuntao Zou[3]

Received: 3 September 2023 / Accepted: 28 January 2024

Abstract

In this study, we aimed to address the primary challenges encountered in industrial integrated circuit (IC) surface defect detection, particularly focusing on the imbalance in information density arising from difficulties in data sample collection. To this end, we have developed a new hybrid architecture model for IC surface defect detection (SDDM), based on ResNet and Vision Transformer (ViT). The core innovation of SDDM lies in the integration of the concepts of image information density and dataset information density, effectively identifying and processing areas of high information density through multi-channel image segmentation techniques. The convolution operations performed within each patch of the model help to precisely capture positional information, thereby meticulously differentiating the complex details on the surface defect detection of ICs. We optimized the model to make it more suitable for industrial applications, significantly reducing computational and operational costs. The experimental results confirmed that the improved SDDM model achieved an accuracy rate of 98.6% on datasets with uneven information density, effectively enhancing the productivity of IC packaging and testing companies, particularly in handling datasets with imbalanced information density.

Keywords Deep learning · Defect detection · Integrated circuit · Intelligent manufacturing

✉ Yuntao Zou
zouyuntao@hust.edu.cn

Xiaobin Wang
wangxb@dgpt.edu.cn

Shuang Gao
single_312@126.com

Jianlan Guo
guojl@dgpt.edu.cn

Chu Wang
faralleywc@gmail.com

Liping Xiong
xionglp@dgpt.edu.cn

[1] Dongguan Polytechnic, Dongguan 523808, Guangdong, China

[2] School of Aerospace Engineering, Huazhong University of Science and Technology, Wuhan 430070, Hubei, China

[3] School of Computer Science and Technology, Huazhong University of Science and Technology, Wuhan 430070, Hubei, China

1 Introduction

Within the industrial production process, surface defect detection [1] is integral for detecting defective integrated circuit (IC) surface, thereby ensuring the ICs' reliability and rate of qualification. Traditional methods of classifying IC surface defects predominantly involve manual detection [2]. However, this approach heavily relies on the inspector's experience and fails to objectively and rigorously evaluate a multitude of IC defects, thus lacking in detection accuracy and efficiency required to meet the standards of mass production quality control. As a result, machine vision methods [3], characterized by high detection speed, accuracy, application flexibility, and strong anti-interference ability, have gradually supplanted manual detection. Nevertheless, IC surface detection currently faces several primary challenges.

(1) The identification of many types of defects is a challenging task: real-world industrial production encompasses numerous types of IC surface defects, such as surface scratches, cover scratches, pin defects, missing characters, unclear printing, and glue overflow (as depicted in Fig. 1a). This necessitates the model's capa-

Published online: 14 February 2024

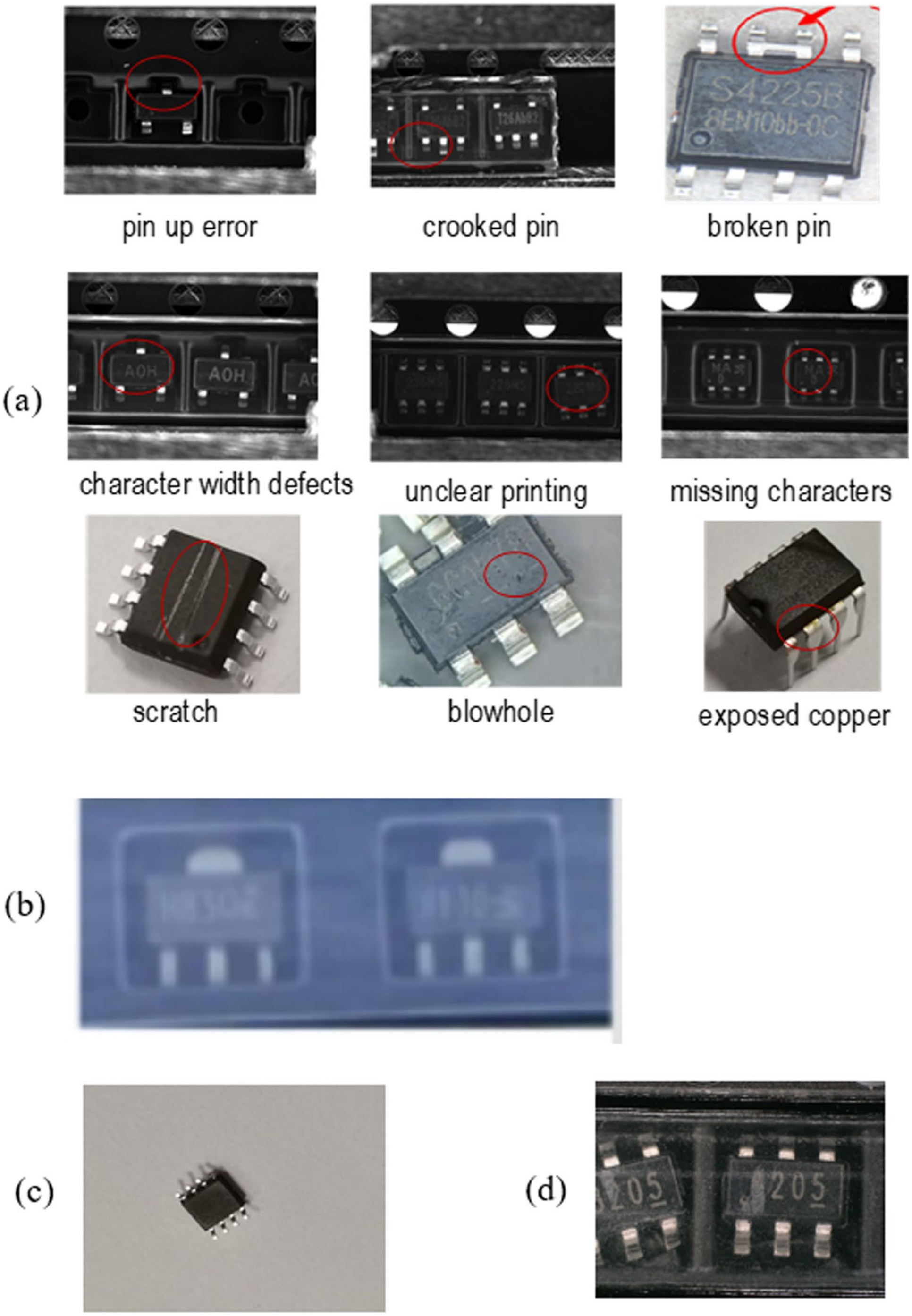

Fig. 1 Issues in IC defect dataset. **a** There are various types of IC surface defects. **b** Character defects require a high recognition rate. **c**, **d** ICs are sourced from different manufacturers and captured in different environments, resulting in dataset diversity

bility to identify a multitude of defects. In addition, the sporadic emergence of novel yet uncommon surface defects demands the model to possess some generalization ability and rapid learning capacity to maintain production efficiency.

(2) The need for improved detection accuracy: as the size of ICs continues to shrink with the progression of integrated circuit technology, the task of identifying defects in IC surface becomes increasingly arduous. Some defects, such as missing characters or non-uniform character width, require high detection accuracy. For instance, the non-uniform character width defect portrayed in Fig. 1b poses a challenge due to the minuscule size of the IC and subtle character width variations. As a general rule, the required accuracy for detecting character width defects is at the scale of tens of microns, thereby necessitating superior detection algorithms.

(3) Dataset collection can be strenuous: modern industrial processes continually optimize, resulting in a decline in defective samples. In a typical dataset, "normal" samples predominantly outnumber "defective" or "abnormal" ones—a phenomenon referred to as "sample imbalance." This disproportion can significantly hinder the application of machine learning-based models.

(4) Large variations in the dataset can create hurdles for model learning: the collection of datasets is a challenging task due to stringent control over IC surface-defective products in manufacturing companies. A sufficient dataset often necessitates image collection from various sources—different factories, data collection equipment, shooting environments, etc.—which results in non-uniform image resolution, size, and information density (as shown in Fig. 1c and d). This lack of uniformity makes the actual dataset greatly differ from standard ones (like ImageNet) and unsuitable for direct processing by machine learning models, constituting another problem that needs to be addressed.

(5) In the actual production process of IC surface defect detection, there is a wide variety of IC specifications and types, with sizes and standards continuously changing. This requires the detection equipment aligned with the production line to adapt to ICs of different specifications and types. In addition, the speed requirements for IC detection are very high, making it impractical for detection equipment (cameras) to continuously adjust focus. Considering cost constraints and the need for speed, standard cameras typically use fixed-focus lenses. To meet the photographic needs of the largest ICs while ensuring the resolution of smaller ICs is not excessively low, the resolution of fixed-focus lenses is set within a specific range. Moreover, incorporating segmentation algorithms in the post-processing would consume substantial resources and fail to meet performance requirements, and enhancing processing capabilities would increase equipment costs. The adaptation to diverse IC specifications and high-speed detection demands, coupled with fixed-focus lens constraints and post-processing limitations, presents significant challenges in the IC surface defect detection process.

In recent years, deep learning has demonstrated significant potential in image classification and detection, with major developments including methods based on convolutional neural networks (CNN) [4] and those utilizing Transformer models [5]. CNNs excel in response speed and accuracy on lightweight datasets, primarily due to their sparse connectivity and weight-sharing characteristics [6]. However, limitations of CNNs include slow processing speeds and inadequacies in acquiring global information [7]. In contrast, Transformer models have been applied to computer vision problems due to their ability to model long-term dependencies and support parallel processing. They are characterized by their simplistic design, capability to handle diverse modalities of data, and exhibit commendable scalability on large-scale networks and datasets. Nevertheless, Transformers exhibit weaker capabilities in handling image-specific inductive biases, necessitating training on extensive datasets, which implies longer training durations and increased computational resource requirements.

In the task of IC surface defect detection, obtaining large datasets of defects is challenging. Using only CNN, it becomes difficult to acquire the global features of images, and a model utilizing only Transformer does not deliver desirable results on smaller datasets. As a solution, this paper proposes an industrial hybrid model SDDM based on both CNN and Transformer, exploiting the strengths of both to learn relevant information on local and global levels. We improved the ViT model to tackle the imbalances present in the IC surface defect dataset. Furthermore, we refined the ViT model to accommodate the uneven information density of the IC defect dataset, enabling the model to handle various types of images obtained from different data sources effectively. When applied to the task of IC surface inspection, the proposed approach demonstrated impressive results.

This paper presents several noteworthy contributions:

(1) This study developed an innovative surface defect detection model (SDDM), ingeniously integrating key features of CNN and Transformers. Tailored for the specific demands of the IC industry work environment, the model underwent performance optimization. This ensured not only higher accuracy in real-world applications but also significantly improved detection efficiency. To address challenges posed by the diversity of image sources, we established a specialized dataset and introduced concepts of image information density and dataset information density variation. These metrics quantitatively describe the diversity within images and datasets.

(2) To address the challenges posed by the diversity of image sources, we have established a dedicated dataset and introduced the concepts of image information density and dataset information density variability. These metrics quantitatively depict the diversity within an image and a dataset. Building upon these measures, we present an enhanced multi-channel image segmentation approach that effectively boosts the detection efficiency of our model.

(3) By implementing model compression relative to data size and computational power demands, we put forth a more compact model that yields impressive results. This compression not only elevates detection efficiency but also lessens hardware demands, thereby enhancing the practical applicability of the proposed model in real-world scenarios.

The remainder of the paper is structured as follows: Sect. 2 provides an overview of the existing work related to surface defect detection and the application of deep learning models in this area. Our proposed model is detailed in Sect. 3, while the experimental results are discussed in Sect. 4. In Sect. 5, we analyzed the experimental results, identifying the deficiencies of our model and outlining the subsequent improvements to be undertaken. The final section outlines the significant contributions of this research.

2 Related Work

2.1 Surface Defect Detection

Surface defect detection involves identifying irregularities such as scratches, flaws, and foreign object interference on sample surfaces [8]. Traditionally, feature-driven machine vision algorithms have been widely used for such tasks. For instance, in 1995, Bennett and Hoy proposed a method that combined multiple comparison detections to identify coplanar pins in IC packaging [9]. In 2010, Liu H and Zhou W introduced a two-dimensional wavelet transform feature extraction algorithm for wafer defect detection [10]. In 2011, D Gnieser and Tutsch R utilized optical detection principles to identify defects and cracks in BGA packaged ICs, and in the same year, Lu et al. employed an improved median filter and Fourier transform for solder joint defect recognition [11]. In 2012, Kaitwanidvilai S proposed a wavelet function transform algorithm for IC pin detection and counting [12], while Berges et al. processed the surface images of MOSFET ICs, effectively measuring pinhole numbers by contrasting the pinhole areas with metal regions [13].

With the rise of deep learning technology, its application in defect detection has also significantly expanded. Ding R et al. proposed a multi-layer deep feature fusion method for calculating the similarity between templates and defective circuit boards, enhancing detection performance [8, 14]. Luan C et al. developed a cross-category defect detection dual-layer neural network that does not require retraining [8]. Anvar A et al. introduced the ShuffleDefectNet, a deep learning-based defect detection system for identifying surface defects on steel strips [15]. Finally, Hu B et al. introduced a new PCB defect detection method based on Faster RCNN, constructing a new network for better detection of small PCB defects [16]. These studies demonstrate the efficient performance of deep learning in the realm of defect detection.

2.2 CNN-Based Model

In 1962, Hubel and Wiesel first introduced the concept of the “receptive field,” laying the foundation for artificial neural networks [17]. In 1980, Fukushima proposed the Neocognitron, considered the prototype of convolutional neural networks (CNNs) [18]. In 1989, LeCun invented CNNs and applied them to handwritten character recognition [19]. In 1998, LeCun introduced the classic LeNet-5 network model, further enhancing the accuracy of handwritten character recognition [20]. In the field of defect detection, CNNs are often combined with other models. Gao X. et al. designed an FCN-based deep learning model for tunnel defect detection [21]. Xiao L. et al. developed the Image Pyramidal Convolutional Neural Network (IPCNN) for surface defect detection [22]. Al-Tam and others proposed a hybrid model combining residual convolutional networks and Transformers for breast cancer auxiliary diagnosis [23]. Jiaxuan Chen et al. applied CNNs and Transformers to multimodal image matching [24], and Guangwei Gao et al. proposed a cooperative network for facial super-resolution recognition [25]. However, hybrid models of CNNs and Transformers have not yet been applied in the field of IC surface defect detection.

2.3 Vision Transformer (ViT)

The Transformer was initially applied in natural language processing (NLP) tasks, first introduced by A. Vaswani et al. for machine translation and English constituent parsing tasks [5]. J. Devlin and others proposed the Bert model based on Transformer, achieving outstanding results in multiple NLP tasks [26]. The GPT model by T. B. Brown et al. also showed remarkable performance [27]. In the field of computer vision, Transformers have been applied to tasks such as object detection, segmentation, and video understanding, achieving optimal performance [28–30]. Vision Transformer (ViT), specifically designed for image classification, utilizes global self-attention mechanisms to model relationships between feature sequences [31]. Recently, researchers have shifted focus towards enhancing local information modeling capabilities, as seen in Transformer-in-Transformer (TNT), Swin Transformer, and regionViT [32–34]. ViT has been applied in various domains, including industrial visual inspection and damage detection in aircraft engine blades [35, 36]. In contrast, CNNs achieve image feature extraction through stacking convolutional layers but face issues of large computational demand and gradient vanishing. Networks based on Transformer, like ViT, although powerful in feature extraction, require training on large-scale datasets. CNN-based models, with their inherent inductive biases, can learn more effectively with less data.

3 Methods

3.1 Model Architecture

The architecture of the proposed IC surface defect detection model is illustrated in Fig. 2. The model primarily consists of the following components: dataset processing, ResNet and ViT feature extraction, decision-level fusion, and final classification.

(1) Data Processing

In the actual industrial production, due to the improvement of technology and process, the proportion of "normal" sample data in the data set is larger, while the amount of "defective" or "abnormal" sample data is small. This leads to the problem of data imbalance. When training deep learning models, it is often required to have a balanced number of samples of each category in the sample set. When unbalanced samples are applied to a supervised learning task, the algorithm will focus more on the categories with large amounts of data and underestimate the categories with small amounts of data. This can cause the model to fail to learn features that appear to be rarely anomalous and thus fail to detect such anomalies. Therefore, we need to process the dataset accordingly. The whole process of dataset processing includes data collection, data cleaning, image labeling, image augmentation and image resizing.

(2) Feature Extraction via ResNet and Improved ViT

After data processing, images are subjected to separate training via the ResNet50 model and an improved variant of the Vision Transformer (ViT) model. The unique strengths of these models—the proficiency of ResNet50 in learning local image features and the improved ViT's adeptness at capturing global features—allow for the distinct extraction of features from IC surface defect data. Considering the uneven distribution of image information density in actual IC surface defect detection tasks and acknowledging that a standard ViT fails to meet the engineering requirements of the task, several improvements have been applied to the standard ViT model. These enhancements aim to boost detection accuracy and speed.

(3) Decision-Level Feature Fusion

Upon ResNet50 and Improved ViT individually extracting features from images, these features are combined in

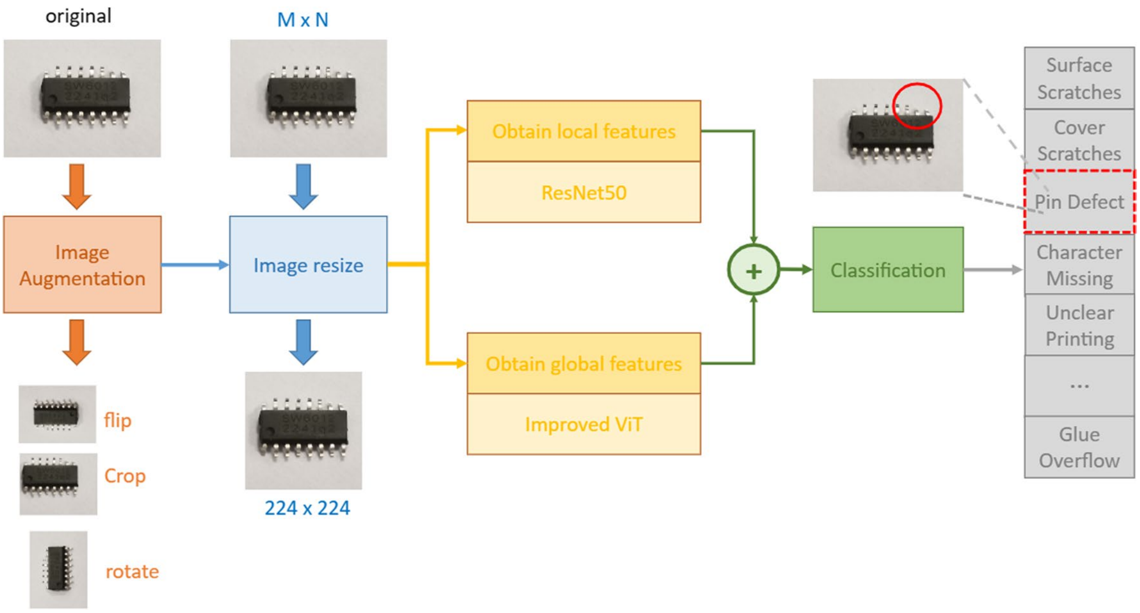

Fig. 2 Our model architecture. The model primarily consists of the following components: dataset processing, ResNet and ViT feature extraction, decision-level fusion, and final classification

a superimposition strategy. This fusion of features ensures that the final attribute set contains both detailed and broad, global features, thereby aiding subsequent classification efforts.

(4) Final Classification

The classification of defect categories is based on superimposed features. The central philosophy of the proposed model involves the separate extraction of local and global features via the ResNet50-ViT amalgamation, followed by the fusion of these feature sets to enhance classification accuracy.

3.2 Image Information Density

In the actual production process of IC surface defect detection, the multitude of IC specifications and types are constantly changing, and differences in images arise from various shooting scenarios, as shown in the Fig. 3. To assess the quality of images and datasets, we have adopted image information density as a measurement metric.

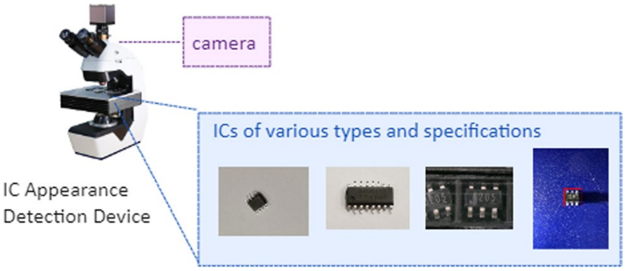

Fig. 3 In the actual production process of IC surface defect detection, the multitude of IC specifications and types are constantly changing

The area of the image is defined as:$S_{\text{img}} = |T_l| \times |T_w|$, and the area of the IC within the image is given by: $S_{ic} = |I_l|_2 \times |I_w|_2$, where $|I_l|_2 = \sqrt{(y_2 - y_1)^2 + (x_2 - x_1)^2}$, and $|I_w|_2 = \sqrt{(y_3 - y_1)^2 + (x_3 - x_1)^2}$. Thus, the formula for information density can be defined as: $\eta = \frac{|I_l|_2 \times |I_w|_2}{|T_l| \times |T_w|}$.

According to the above formulas, we can calculate the information densities of Fig. 4b and c as follows: the information density of the Fig. 4b is $\eta_b = 0.35$ while the information density of the Fig. 4c is $\eta_c = 0.04$. Evidently, the image information density of (b) is higher than (c).

Assuming that the data set $\mathcal{D}$ comprises of n images, with $d_1, d_2, \ldots, d_n$ representing different samples in $\mathcal{D}$ and $\eta_1, \eta_2, \ldots, \eta_n$ representing the corresponding image information destinies. The average image information density is denoted as $\overline{\eta}$. Subsequently, we utilize v to characterize the differences of dataset information density across various data sets, thus

$$v = \frac{1}{n} \sum_{i=1}^{n} (\eta_i - \overline{\eta})^2$$

If the images within a dataset are all captured by the same entity and under identical conditions, v will be relatively small; conversely, v will be large. A larger v implies uneven information density across the dataset.

3.3 Improvement on the ViT Structure

(1) Multi-channel Segmentation

The conventional ViT's structure is illustrated in Fig. 5. Its core functionality involves segmenting images into patches according to a $P \times P$ matrix, fusing the segmented

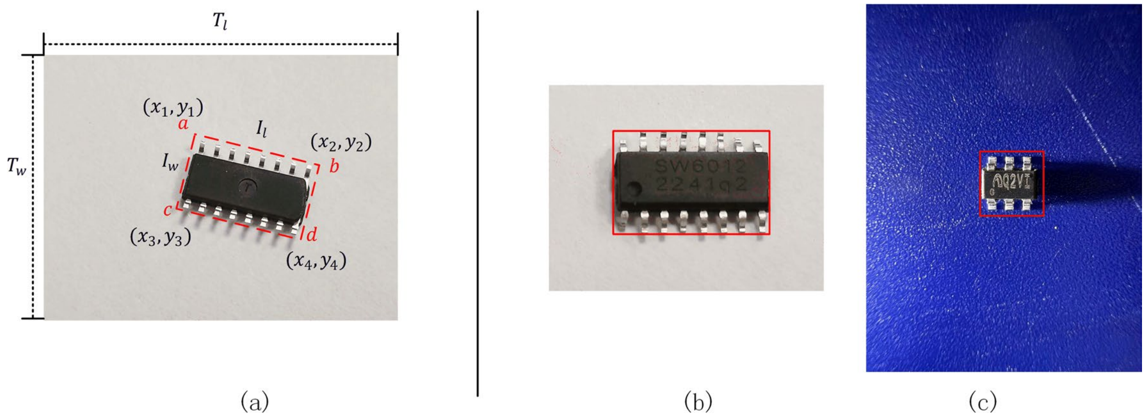

Fig. 4 Image information density. **a** Calculation method of information density. **b**, **c** Images with different information densities. The image information density of (**b**) is higher than (**c**)

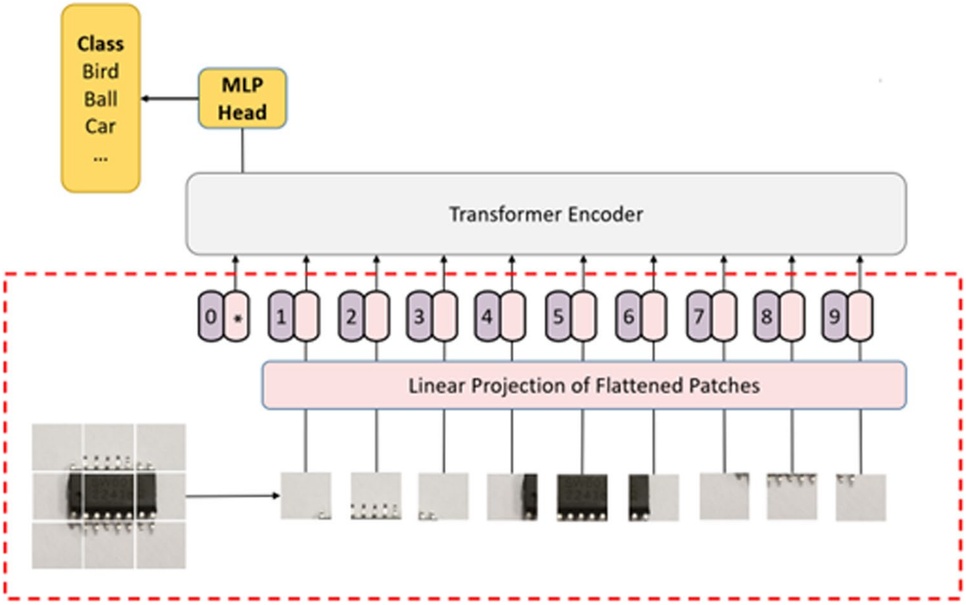

Fig. 5 Structure of the traditional ViT. Traditional ViT structure. In the traditional ViT structure, a single size $P \times P$ matrix is used to segment the image

patches and position embedding via linear projection, followed by a direct application of the Transformer model. However, the direct utilization of the traditional ViT model for IC surface defect detection presents certain challenges. This is because the ViT model's training employs datasets such as ImageNet and Google's Open Images, which typically have more uniform image sizes and balanced image information densities.

Contrarily, for our task, gathering an IC surface defect dataset is challenging due to the various image sources. This makes it hard to ensure that all images within the IC Surface Defect Dataset possess similar image information densities, as observed in Fig. 6a with high information density and Fig. 6b with low image information density. The application of a fixed $P \times P$ matrix set for segmentation can be problematic. For instance, employing the same 3×3 matrix for segmentation could work for Fig. 6a but not for Fig. 6b, where the method segregates the IC image with higher image information density into one patch, leaving the other background images with lesser information density in other patches. This approach is not conducive to image feature learning. Compared to Fig. 6b, c offers a more rational image segmentation.

To resolve this issue, we have designed multiple segmentation channels, i.e., providing $P_1, P_2, ..., P_i$ multiple segmentation methods to segment the image, as shown in Fig. 7. This ensures that areas with high information density can be segmented into different patches. While more segmentation methods may offer more feature learning opportunities, they also increase computational requirements. In the specific process of IC surface defect detection, companies must factor in cost-efficiency. Hence, a balance between limited computational resources and more segmentation methods is necessary. Experimental findings suggest that two sets of segmentation methods, $P = 16$ and $P = 32$, strike a good balance for our dataset. Thus, we have designed the

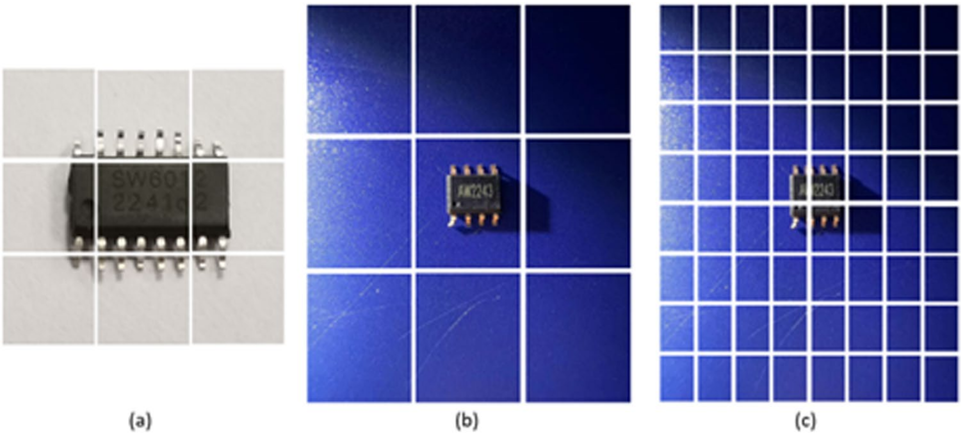

Fig. 6 Different ways to segment images. **a**, **b** exhibit different image information densities. If a 3×3 matrix is used to segment the images, **a** can be effectively segmented, while (**b**) is not properly segmented. In (**c**), an 8×8 matrix is used to segment the image, which allows for effective segmentation compared to (**b**)

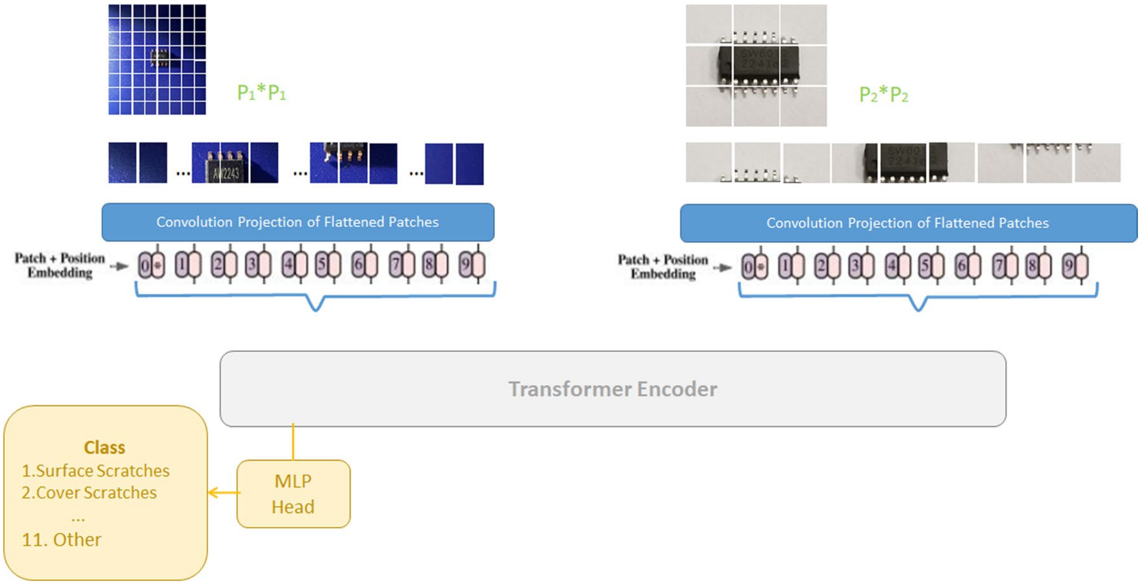
P₁*P₁
P₂*P₂
Convolution Projection of Flattened Patches
Convolution Projection of Flattened Patches
Patch + Position Embedding
Patch + Position Embedding
0* 1 2 3 4 5 6 7 8 9
0* 1 2 3 4 5 6 7 8 9
Transformer Encoder
Class
1.Surface Scratches
2.Cover Scratches
...
11. Other
MLP Head

Fig. 7 We improved the traditional ViT model by introducing multi-channel image segmentation and performing convolution within each patch, enhancing the model's ability to detect subtle IC defects caused by uneven information density in the images

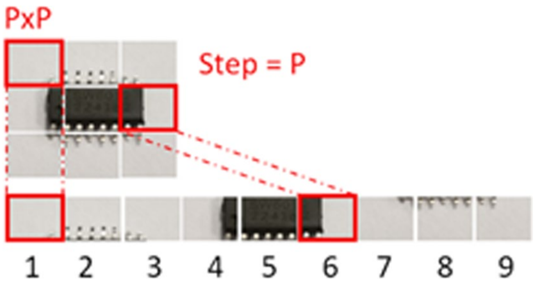

Fig. 8 Convolution within each patch. We have improved the traditional ViT model by introducing the process of performing convolution within each patch

model as a dual-channel model. However, in other scenarios where computational resources are not a constraint, and the problem of unbalanced image information density exists, a multi-channel model could be designed to enhance accuracy.

(2) Executing Convolution Within Each Patch

In its standard operation, ViT, following image segmentation, converts the image into a vector via Patch and Position Embedding, thereby integrating the positional information into the patches. This approach effectively addresses the lack of positional information in self-attention and handles positional information between patches. However, the positional information within each patch is lost since the patch image information is directly converted from 2 to 1D, as illustrated in Fig. 8. While information between patches (e.g., patches 1, 2, 3, …, 9) is retained, the unidentified internal information within each patch (e.g., patch 1 or patch 5) is lost. Although the positional information between patch blocks can somewhat compensate, it still negatively impacts the convergence efficiency. In the IC detection task, where the image information density is uneven, this loss of internal positional information within patches may affect the results.

In fact, during the ViT task segmentation, it can be perceived as applying a convolution-like operation with a field of view of $P \times P$ and a stride size of P for the segmentation, as shown in Fig. 8. If we introduce a convolution operation at this stage, namely executing a convolution with a field of view of $P \times P$ and a stride length of P, we can secure the image positional information within the patch while segmenting the image.

The algorithm is illustrated in Algorithm 1. Experimentally, this modification has proven beneficial for IC detection, outperforming simple linear transformations.

Moreover, we discovered that performing convolution inside the patch not only provides the internal positional information of the patch but also yields another advantage: post-convolution, the size of the patch is compressed while the token remains the same. Taking an original patch size of 16×16 as an example, the initial size is $16 \times 16 \times 3 = 768$. After executing the convolution, the size becomes $5 \times 5 \times 3 = 75$. We employ the compression ratio to quantify the degree of image compression:

$$\text{compression ratio} = \frac{\text{size of Conv}}{\text{size of original}} = 9.76\%.$$

That is, the size of the vector post-convolution is only 9.76% of the original, yielding a compression ratio of about 1:10. Correspondingly, with Token = 196, this results in a memory reduction by $(16 \times 16 \times 3) \times 196 - (5 \times 5 \times 3) \times 196 = 13528$. This considerable saving in storage space, along with the simplified computation process, leads to an increase in computation speed and a reduction in hardware requirements, which is of significant importance for engineering applications.

Algorithm 1 Algorithm 1

```
Input: X(size16*16)
Output: X(size5*5)
Conv2d(3, kernel size = 7, stride = 2, padding = 3)
X(size8*8)
Relu
MaxPool2d(kernel_size = 2, stride = 2, padding = 1)
X(size = 5*5)
Flatten
```

(3) Dual-Channel Vector Unification

After modifying the ViT model to include dual-channels, one challenge we faced was the variation in vector sizes. The most straightforward and common solution is to connect a separate Transformer Encoder to each channel. However, in the task of IC surface defect detection, resource constraints such as storage and computation necessitate a more cost-effective solution. Our solution is to unify the vectors from both channels, enabling the use of just one set of Transformers to handle vectors from different channels, as depicted in Fig. 7.

In the image resizing process, we adapt the image to a 224×224 size to simplify further processing. Specifically, we have

$$F_{\text{resize}}(X_{n\times m}) = X_{224\times 224}$$

We set the first segmentation size P_1 to 16 and the second segmentation size P_2 to 32, i.e., P_1=16 and P_2 =32. Following convolution, we have

$$F_{\text{Conv}}(P_1.\,\text{size}.16 \times 16) = P_1.\,\text{size}.5 \times 5, \text{ and}$$

$$F_{\text{Conv}}(P_2.\,\text{size}.32 \times 32) = P_2.\,\text{size}.16 \times 16.$$

$$\text{size}.\,P_1 = 5 \times 5 \times 3 = 75,$$

$$\text{size}.\,P_2 = 16 \times 16 \times 3 = 768.$$

After image segmentation, the image vectors are converted from 2 dimensional vectors to 1 dimensional vectors, the vector size of Patch1 is 75 and the vector size of Patch2 is 768:

$$\text{No. } P_1 = \frac{224 \times 224}{16 \times 16} = 196, \text{ and}$$
$$\text{No. } P_2 = (224 \times 224)/(32 \times 32) = 49.$$

The token number of the segmented image Patch1 is 196, and the token number of the segmented image Patch2 is 49.

For a 16 × 16 patch, we use a 75 × 75 matrix for linear changes. In this way, ViT still receives a 196 × 75 Matrix:

$$\text{Patch}_{16}(196 \times 75) \times (75 \times 75) = 196 \times 75$$

In order to allow the 32 × 32 patch to use the same Transformer Encoder, we use a 768 × 75 matrix to convert the size of P_2 to 49 × 75, that is

$$\text{Patch}_{32}(49 \times 768) \times (768 \times 75) = 49 \times 75.$$

This can be directly processed by the same Transformer Encoder.

3.4 Model Advantage

Our proposed solution introduces an IC surface defect detection model, SDDM, which is a fusion of ResNet50 and improved ViT. The model enhances the traditional ViT to meet the specific requirements of the task. This model harnesses the single strengths of ResNet50 and ViT to extract local and global features, which are then amalgamated to categorize defect types, thereby enhancing classification accuracy. Moreover, to address the uneven distribution of image information density in IC surface defect detection, we propose a dual-channel segmentation strategy to boost classification accuracy. By incorporating convolution within the patch, we are able to compress the model, thereby reducing costs and aligning more closely with engineering requirements.

The key benefits of our model are as follows:

(1) We introduce the SDDM, a ResNet50 and Transformer hybrid model, which is able to acquire image features from both local and global perspectives. This Dual-channel facilitates superior extraction of the detailed features from various classes of IC surface defects, thereby enhancing defect recognition efficacy.

(2) We adapt the traditional ViT to address the challenge of uneven image information density in the IC surface defect detection dataset. We propose a Dual-channel Transformer, whereby varying image segmentation sizes ensure effective segmentation of high informa-

tion density regions within the images, further boosting recognition efficiency.

(3) By implementing convolution within the patch, the model not only captures the detailed nuances of IC surface defects more effectively, but also compresses the model. This compression reduces computational and operational costs, making the model more applicable for engineering contexts.

4 Experiment and Results

In this section, we initially discuss the process of dataset preparation. Following this, our proposed methodology is evaluated on the ImageNet dataset, aimed at verifying the effectiveness of the enhancements made to the ViT technique. Lastly, we validate our proposed approach using an IC defect dataset and draw comparisons with contemporary leading methodologies, demonstrating the superiority of our approach.

4.1 Dataset Processing

The IC surface defect dataset is employed for the training and testing of our SDDM model. Initially, images from various factories were collected, followed by data cleansing and image labeling, culminating in the creation of the IC surface defect dataset. This dataset contains a total of 2043 original images, of which 1501 are positive samples without defects, and 542 are negative samples with defects. The negative samples encompass a range of 11 categories including surface scratches, abnormal cover film scratches, pin problems, and missing characters, as illustrated in Table 1.

Upon analysis, we find an approximate 3:1 ratio of positive to negative samples in our dataset, with negative samples representing a smaller fraction. Furthermore, an imbalance also exists among the negative samples themselves. For instance, the sample count for defects such as Copper leakage and Air hole is particularly low, creating a 214:1 ratio when compared with positive samples. Such dataset imbalance significantly impacts the effectiveness of training. To enhance anomaly recognition, it is essential to balance the data, which necessitates data augmentation for negative samples within the dataset.

Table 1 Original negative sample

No	Model number	Quantity
1	Surface scratches	12
2	Cover film scratches abnormal	22
3	Pin problem	100
4	Missing characters	11
5	Inconsistent character width	142
6	Blurred characters	139
7	Copper leakage	7
8	Air holes	7
9	IC defect	13
10	Glue overflow	9
11	Other	77

(1) Data Augmentation

We employ five distinct data augmentation techniques to expand our data. These methods include image flip (horizontal and vertical), rotation, scaling, cropping, and translation. As depicted in Fig. 9, the defective dataset expanded from an original 542 to 7588 instances.

The flip operation incorporates both horizontal and vertical flipping. The rotation technique involves two rotations of 90 and 180 degrees. The scaling technique is designed to increase the image size by scaling inwards, and a portion equivalent to the original image size is segmented from the newly enlarged image. This process is performed twice. Cropping involves randomly sampling a portion from the original image and resizing this part to match the original image's size; this operation is performed four times. Finally, the translation technique includes four panning operations: left, right, up, and down.

In the enhanced dataset, we have 1501 defect-free samples and 7588 defective samples, as shown in Table 2. The dataset is further segmented into a training and test set and comparative set in an 8:2 ratio, utilizing 7210 images for training and set and 1879 for model comparison.

(2) Image Resizing

The images comprising our dataset were sourced from various companies and different production lines, leading to significant variations in image size, layout, and a multitude of other characteristics. The original dataset includes a diverse array of image resolutions, with the four primary resolutions being 512 × 480, 1440 × 1080, 4608 × 3456, and 1276 × 1702 pixels. The process of data augmentation further generates a wider range of image sizes.

Subsequent Transformer and CNN frameworks impose strict size requirements on incoming images. To fulfill the data prerequisites of these models, we implement a resizing operation post data augmentation. To facilitate superior training outcomes, all images within the dataset are resized to 224 × 224 pixels, matching the standard image size in the ImageNet dataset.

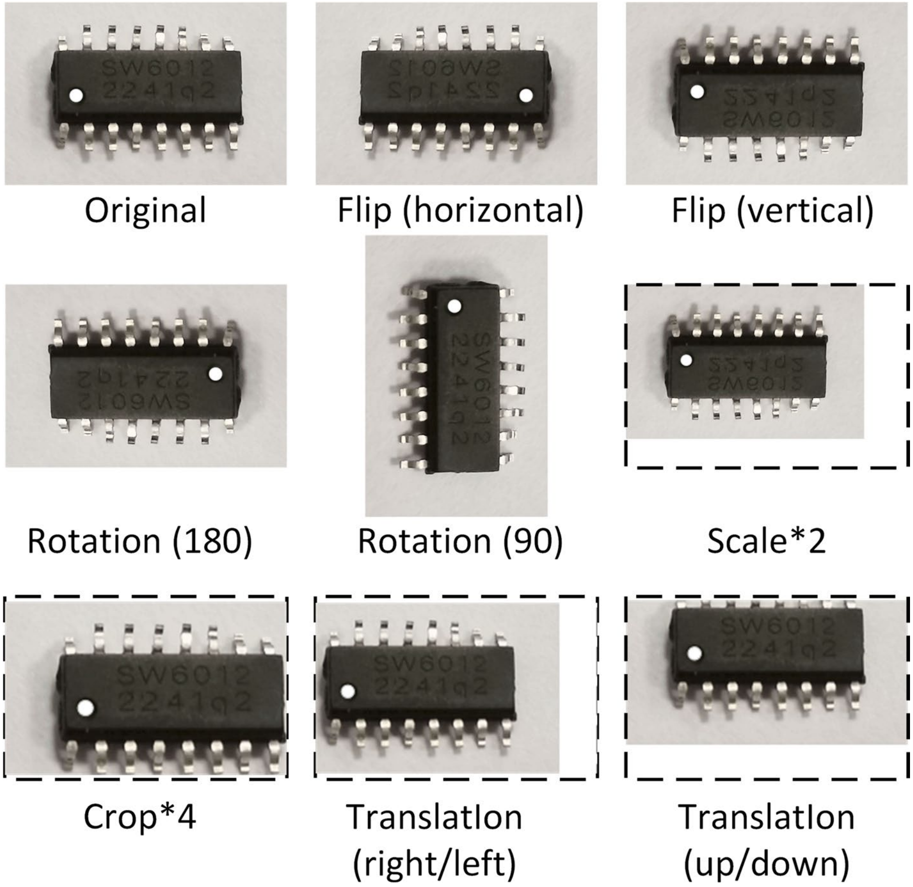

Fig. 9 Data augmentation method. The data augmentation methods used in our study include Flip, Rotation, Scale, Crop, Translation, and others

Table 2 Augmented negative sample

No	Model number	Quantity
1	Surface scratches	720
2	Cover film scratches abnormal	660
3	Pin problem	700
4	Missing characters	660
5	Inconsistent character width	705
6	Blurred characters	695
7	Copper leakage	700
8	Air holes	700
9	IC defect	650
10	Glue overflow	630
11	Other	768

4.2 Performance Evaluation

(1) Experiment Procedures

First, to validate the effectiveness of the proposed model, a detailed decomposition and comparative analysis of its main components were conducted. The results of these comparative experiments are presented in Sect. 4.3.1 of this paper. The findings indicate that the optimized segments of our project exhibit performance enhancements over the original model. Section 4.3.2 details the outcomes of the complete model training. A comparison with some of the latest models reveals that the Superiority of the SDDM becomes increasingly apparent with the rise in image information density.

(2) Experimental Settings

The loss functions for both models need to be determined by

$$\text{Loss} = \sum_{i=1}^{i=n} |W_i \times \text{Loss}_{\text{model}_i}|/n.$$

Let $W_1 = a_{\text{ResNet50}}$, $\text{Loss}_1 = L_{\text{ResNet50}}$, $W_2 = a_{\text{ImprovedViT}}$, $Loss_2 = L_{\text{ImprovedViT}}$, we have

$$\text{Loss} = a_{\text{ResNet50}} \times L_{\text{ResNet50}} + a_{\text{ImprovedViT}} \times L_{\text{ImprovedViT}}.$$

Here, L_{ResNet50} and $L_{\text{ImprovedViT}}$ represent loss function of ResNet50 and improved ViT, a_{ResNet50} and $a_{\text{ImprovedViT}}$

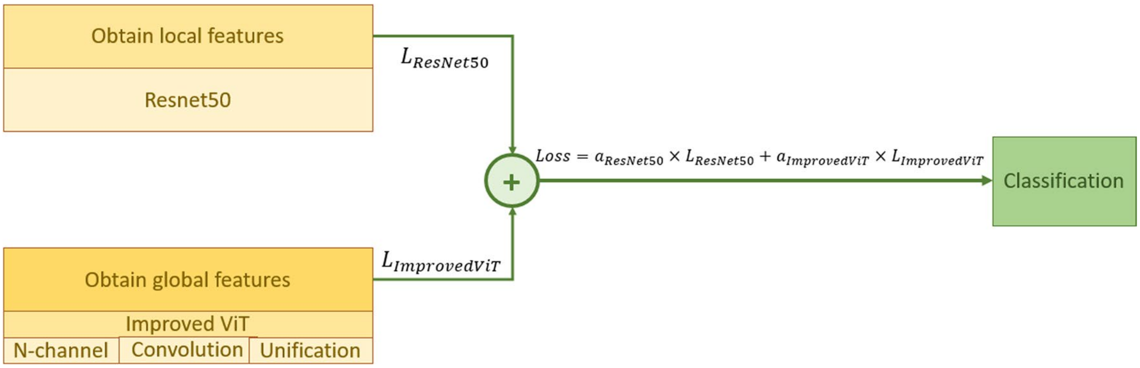

Fig. 10 The calculation of the loss function is Loss $= a_{\text{ResNet50}} \times L_{\text{ResNet50}} + a_{\text{ViT}} \times L_{\text{ImprovedViT}}$, where L_{ResNet50} and $L_{\text{ImprovedViT}}$ represent the loss function of ResNet50 and improved ViT, a_{ResNet50} and a_{ViT} represent the model weight coefficients, respectively

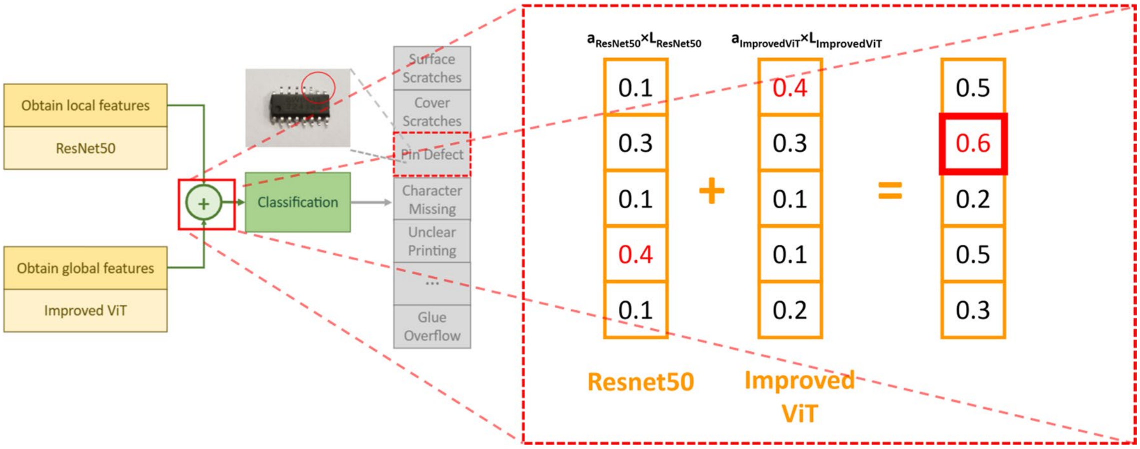

Fig. 11 The loss function is calculated by assigning equal weights $a_{\text{ResNet50}} = a_{\text{ImprovedViT}} = 1/2$, to the hybrid model composed of ResNet50 and ViT

represent the model weight coefficients, which can be trained through linear regression or other suitable techniques, as shown in Fig. 10. In this study, for simplicity, we assume equal weights for both models, i.e., $a_{\text{ResNet50}} = a_{\text{ImprovedViT}}$ =1/2. These weights can be optimized at a later stage to improve overall performance. The final loss function is computed as: Loss $= |L_{\text{ResNet50}} + L_{\text{ImprovedViT}}|/2$, as depicted in Fig. 11.

The learning rate in this methodology is initialized at 0.001 and undergoes decay via a cosine learning rate. We undertake 200 epochs of training employing the Adam optimizer, given its capability to facilitate faster network convergence and more effective parameter adjustment. In the ViT model, we employ GELU as the activation function in conjunction with the L2 regularization. In the ResNet50, we use the rectified linear unit (ReLu) along with the L2 regularization. To avoid bias in the results, the same parameter configurations are applied across all comparisons. We implement our method using the Pytorch toolbox, with the momentum coefficient of the network set to 0.9. Our experiments are conducted on a server equipped with the following key configurations: dual Xeon 10-core processors, dual 3090 graphic cards with 24 G RAM (48 G in total), 4 sets of Samsung 32 G RAM, a 500 G SSD, and a 4 T data disk.

4.3 Experimental Results

4.3.1 Comparative Analysis of Model Components and the Benefits of Improvement

First, a comparative analysis is conducted between our proposed hybrid model and the single models of ResNet50 and ViT. This analysis aims to evaluate whether the hybrid approach demonstrates superior performance compared to the single models. Subsequently, the improved ViT model, which incorporates internal patch-wise convolution, is compared with the standard ViT model. This comparison is intended to assess if the incorporation of internal convolution in the improved ViT model results in higher recognition accuracy. Lastly, we juxtapose the improved ViT model, employing a dual-channel approach with two different patch sizes for image segmentation, against the standard single-patch-size ViT model. The purpose of this comparison is to examine whether the dual-channel approach in the improved ViT model can potentially enhance the model's recognition precision.

To ensure the fairness of the experiments, all models in this section were trained exclusively using the ImageNet dataset and then tested using the IC Surface Defects Testing Dataset.

To assess the model's performance, the metric of Accuracy is employed. Accuracy is calculated using the following formula and reflects the model's ability to correctly classify ICs with and without surface defects:

$$\text{Accuracy(acc)} = \frac{\text{Number of Correct Predictions (right)}}{\text{Total Number of Predictions (all)}}.$$

Here, 'Number of Correct Predictions' (right) includes True Positives (TP) and True Negatives (TN), which are the counts of correctly predicted positive and negative classes, respectively. 'Total Number of Predictions' (all) comprises all True Positives, True Negatives, False Positives (FP), and False Negatives (FN).

(1) Comparative Analysis of the Hybrid Model and Single Models

To assess the performance of our proposed hybrid model, we executed comparisons against the single ResNet50 and ViT models. It should be noted that in these comparisons, the original ViT, not the improved version, is utilized within the hybrid model to preclude extraneous interferences. The experiment result is shown in Table 3.

From the experiments, it is observed that:

ResNet50 achieves an accuracy rate of 69.71%, outperforming the ViT model, which attains only 66.45% accuracy. This represents a 3.26% improvement, highlighting ResNet50's superiority over ViT in small-scale datasets.

The hybrid model records an accuracy of 72.51%, exceeding the single ResNet50 and ViT models by 2.8% and 6.06%, respectively. This demonstrates its superior performance.

Table 3 Comparison of the hybrid model with single ResNet50 and ViT models. The hybrid model showcases a higher accuracy rate compared to the single models on IC surface defects dataset

Model	Acc (%)
ResNet50	69.71
ViT	66.45
ResNet + ViT hybrid model	72.51

To summarize, the hybrid model exhibits a higher accuracy rate compared to the single models, ResNet50 and ViT. However, while the hybrid model shows some improvement over the single models, the increase is not substantial. Overall, the performance of the models on the IC dataset compared to their performance on ImageNet shows a significant gap, indicating that the results are still not optimal. This suggests that merely making macro-level improvements or combining frameworks is not sufficient to address issues of uneven information density effectively. Therefore, further optimization of the model's details is necessary to achieve better results.

(2) Comparison of Improved ViT After Performing Convolution and Standard ViT

To evaluate the improvement of model performance after performing convolution inside the 16 × 16 patch, we compared the ViT after performing convolution with the standard ViT model, and the experimental results are shown in Table 4, with comparison of ViT for performing convolution with standard ViT.

With the above experiments, we have the following findings: the accuracy of the model after performing convolution inside the 16 × 16 patch is 69.18%, which is 2.73% higher than the standard ViT of 66.45%, and the performance of the model has been improved.

In addition, upon further observation and analysis, it has been discovered that the Transformer's multi-headed attention mechanism is more adept at detecting 'pinning errors', a type of inconsistency in detail errors. This level of nuanced error detection is challenging for traditional CNNs, which are limited to only conducting local analysis.

(3) Comparison of Multi-channel ViT and Standard ViT

To assess the effect of our enhanced multi-channel ViT on model performance, we conducted a comparison between

Table 4 Comparison of ViT for performing convolution with standard ViT. The accuracy of the model after performing convolution inside the 16 × 16 patch is 2.73% higher than the standard ViT

Model	Acc (%)
ViT	66.45
ViT + Conv	69.18

Table 5 Comparison of multi-channel ViT with standard ViT. The accuracy rate after implementing dual-channel ViT for image segmentation is 67.83%, representing a 1.38% increase over the standard ViT's 66.45%

Model	Acc (%)
ViT	66.45
2-channel-ViT	67.83

a dual-channel ViT (using 16 × 16 and 32 × 32 patches) and the standard ViT. The experimental results are presented in Table 5, comparing the performance of the multi-channel ViT and standard ViT.

From the experimental results, we observe that the accuracy rate after implementing dual-channel ViT for image segmentation is 67.83%, representing a 1.38% increase over the standard ViT's 66.45%, thereby indicating a certain degree of performance enhancement.

(4) Short Summary

- Through our experiments, it is observed that the proposed ResNet + ViT hybrid model attains an accuracy of 72.51% on the IC surface defects dataset, which represents a gain of 2.8% and 6.06% compared to the standard ResNet50 and ViT models when used independently, thereby corroborating the efficacy of our proposed hybrid model.
- Upon implementation, we noted that the model's accuracy post-execution of the convolution within the 16 × 16 patch reached 69.18%, marking a 2.73% improvement over the standard ViT's 66.45%, thereby illustrating an enhancement in the model's performance.
- From the experimental outcomes, it is inferred that the accuracy after the application of the Dual-channel ViT for image segmentation is 67.83%, exhibiting a 1.38% rise compared to the standard ViT's 66.45%, indicative of a notable improvement in accuracy.

These findings serve to substantiate that our proposed model effectively enhances the recognition rate.

4.3.2 Model Performance on IC Surface Defects Dataset

(1) Experiments Procedure

In this part, the complete SSDM model was trained using the training set from the IC Surface Defects Dataset. The model was then tested using four different datasets, each characterized by varying information density levels, with respective v values of 0.000189, 0.000544, 0.001841, and 0.008004. It is important to note that the data used for testing had not been employed during the training phase. To ensure fairness, several other contemporary models were also tested under the same datasets. The performance of the SSDM model was then evaluated by comparing its results with those of the other models.

(2) Comparison of Our Model with Current Popular Model

In order to evaluate the performance of the proposed model, we compared it with the current popular models (DINO-DETR, DAB-DETR, DN-Detr, etc.) on the IC surface defect dataset. The experimental results are shown in the Table 6.

From Table 6, the following observations can be made:

(1) When the information density difference v is small, such as when v equals 0.000189, the recognition rates of our proposed model, SSDM, and other currently popular models are not markedly different, as they all demonstrate high accuracy. This can be attributed to the fact that a smaller v indicates minimal differences in the information density among the images within the dataset, therefore, the inherent advantages of our model have not been fully realized.

(2) As v increases, for instance, when v equals 0.008004, the recognition accuracy of other models gradually decreases, whereas our model consistently maintains a high recognition rate. This underlines the robustness and adaptability of our model in handling datasets with significant variations in information density.

Figure 12 shows how our model compares with other models for different information density differences. From the experimental results, the accuracy of our model has been at a relatively high level compared with other models.

Figure 13 illustrates the comparative advantage of our model over the best-performing model among current ones under different degrees of information density differences within the dataset. As depicted in the figure, our model shows superior performance when dealing with datasets characterized by larger discrepancies in information density. This suggests that the larger the information density difference within the dataset, the more our model outperforms

Table 6 Comparison of our model with current popular model in IC surface defect dataset

v	ResNet50 (%)	ViT (%)	DINO-DETR (%)	DAB-DETR (%)	DN-Detr (%)	DEFORME (%)	SDDM (ours) (%)
0.000189	99.20	99.20	99.40	99.20	99.40	99.40	99.60
0.000544	98.25	98.38	98.50	98.88	98.63	98.75	99.13
0.001841	96.90	97.10	97.70	98.00	97.80	98.10	98.80
0.008004	95.67	95.87	96.33	97.00	97.13	97.20	98.60

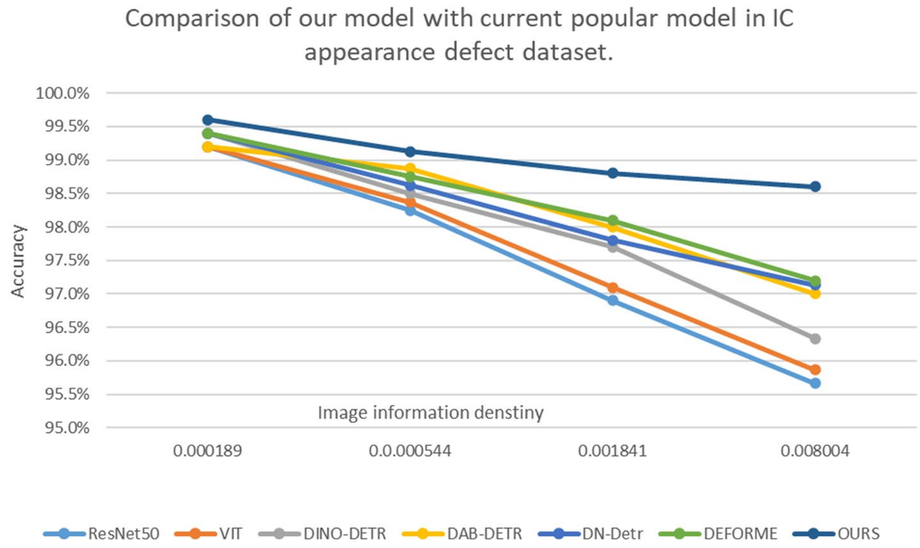

Fig. 12 Comparison of our model with current popular model in IC surface defect dataset

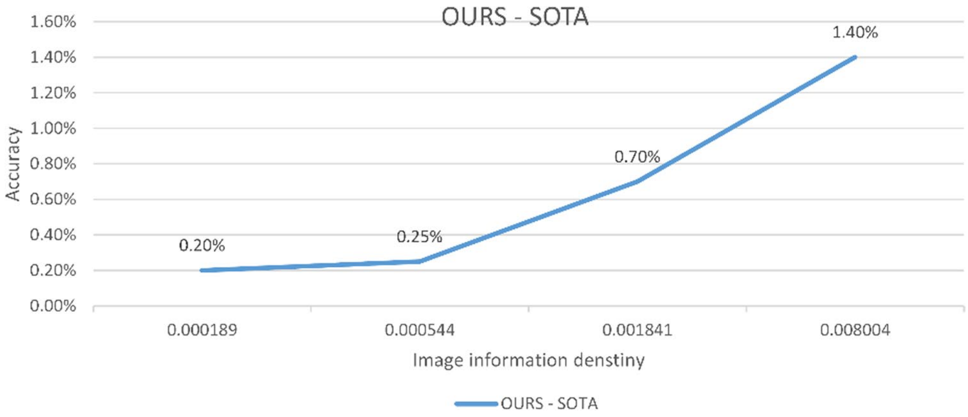

Fig. 13 Comparing our model with the currently best-performing model, we found that our model exhibits greater advantages as the diversity of information density in the dataset increases

others, demonstrating its robustness and adaptability in handling complex and diverse datasets.

(3) Short Summary

Upon comparative assessment with contemporary popular models, our proposed model demonstrates superior performance on the IC surface defects dataset. Its efficacy surpasses that of other prevalent models, and notably, the model's advantage becomes increasingly apparent as the differences in information density within the dataset amplify. This highlights the model's capability to effectively handle complex datasets with varying information densities.

5 Discussion

5.1 Analysis

(1) The integrated model advances detection accuracy beyond that achievable by independent models, a finding consistent with expectations. This is attributed to the amalgamation of the ResNet50's local learning capabilities and ViT's global learning potential within the hybrid model. Consequently, the model is better positioned to extract features of IC surface defects, thereby resulting in superior accuracy.

(2) We theorize the accuracy enhancement post-convolution within the Patch owing to the following reasons: the task of our study revolves around IC surface defect detection, where the disparity between defective samples, as well as between defective and normal samples, predominantly lie in details, thereby distinguishing it from traditional classification tasks. For standard classification tasks such as distinguishing between animals or fruits, discernible features exist both globally and locally. However, within the purview of this project, defect disparities are mainly local and detailed. Thus, conducting convolution within the Patch enables the extraction of more detailed information, enhancing accuracy.

(3) The recognition accuracy saw a modest increase of 1.38% (in the ImageNet dataset) post-execution of dual-channel image segmentation. This restricted improvement is due to the limited computational power provided by the engineering setup, rendering the handling of images with intricate origins via dual-channels challenging. In addition, the improvements seen with dual-channels are primarily applicable to datasets characterized by significant uneven information density; datasets with high-quality images and balanced information distribution may not necessarily witness substantial enhancements.

(4) Our model excels in the IC surface defect dataset due to our consideration of varying information densities across different images within the dataset. By employing dual-channel segmentation, areas of high information density within images can be efficiently segmented. This approach, therefore, proves apt for datasets with complex data sources and aligns with engineering requirements.

5.2 Limitations and Future Improvements

Despite the enhancements in the accuracy of IC surface defect detection achieved through our hybrid model in this study, certain limitations persist within our approach.

First, the volume of our datasets, particularly defective ones, is inadequate, and the overall quality of dataset samples could be improved. In response to these issues, we intend to undertake further data collection efforts. Continual augmentation of both the quantity and quality of data should facilitate more effective feature learning for our model.

Second, due to computational power constraints, we resort to a dual-channel strategy for image segmentation. However, employing fixed size segmentation for images of varying dimensions does not yield optimal results. In future iterations, we could consider an adaptive approach to image segmentation. Moreover, transitioning from a dual-channel to a multi-channel configuration could conserve computational power while producing superior outcomes.

Third, regarding the model's loss function, we have only computed instances where the weight is at 50%. Going forward, we could procure weights through model training. Ideally, the model should dynamically adjust the algorithmic weights according to differing inputs, fostering improved model outcomes.

Lastly, we mainly focused on the accuracy metric to evaluate model performance. Nonetheless, other metrics such as precision, recall, and the F1 score are equally important. For example, precision—the ratio of actual defect-free products among those classified as defect-free—should surpass a specific threshold. Otherwise, defective IC products could enter the market, negatively affecting consumer experience and the company's reputation. Thus, reducing false positives is crucial in model tuning. Similarly, maintaining recall within certain limits is vital to avoid mistakenly discarding defect-free IC products, preventing wastage. Several metrics need to be considered in actual production.

6 Conclusion

In this study, the proposed SDDM model demonstrates significant practicality and effectiveness in the field of IC surface defect detection, especially in dealing with complex image backgrounds characterized by uneven information density. Through comprehensive and rigorous evaluation, our hybrid model has been proven to surpass current mainstream methods in terms of performance, reflecting its reliability and applicability in industrial-level applications. The SDDM model achieved a high recognition accuracy rate of 98.6% on the IC defect detection dataset, significantly enhancing detection accuracy and efficiency, particularly in processing images with varying areas of information density. The results of this study not only showcase the powerful potential of the hybrid model in handling complex image tasks but also provide new perspectives and methodologies for future research and applications in related fields under such uneven data background scenarios.

Acknowledgements This work was supported in part by Dongguan Polytechnic Special funding for the electronic information engineering technology specialty group of the National Double High Program (No. ZXF015), in part by Dongguan Sci-tech Commissioner Program (No. 20221800500492/20221800500552), in part by Songshan lake Sci-tech Commissioner Program (No. 20234420-01KCJ-G), in part by Dongguan Social Science and Technology Development Project (20211800900592), in part by Special fund for Dongguan's Rural Revitalization Strategy in 2021(No. 20211800400102), in part by the Project for the Innovative Platform in Industry and Education Integration for higher vocational college of guangdong province - the innovative platform in industry and education integration of testing technology for intelligent terminal product (No. 2020CJPT014), in part by Dongguan Special Commissioner Project (No. 20211800500182), in part by Guangdong-Dongguan Joint Fund for Basic and Applied Research of Guangdong Province (No. 2020A1515110162) and in part by University Special Fund of Guangdong Provincial Department of Education (No. 2022ZDZX1073).

Funding This study was supported by Dongguan Polytechnic Special funding for the electronic information engineering technology specialty group of the National Double High Program, No. ZXF015, Xiaobin Wang and Shuang Gao, Dongguan Sci-tech Commissioner Program, 20221800500492/20221800500552, Xiaobin Wang and Shuang Gao, Songshan lake Sci-tech Commissioner Program (No. 20234420-01KCJ-G), Shuang Gao, Dongguan Social Science and Technology Development Project (20211800900592), Liping Xiong, Dongguan's Rural Revitalization Strategy, No. 20211800400102, Jianlan Guo, the Project for the Innovative Platform in Industry and Education Integration for higher vocational college of guangdong province - the innovative platform in industry and education integration of testing technology for intelligent terminal product (No. 2020CJPT014), Shuang Gao and Liping Xiong, Dongguan Special Commissioner Project, No. 20211800500182, Jianlan Guo, Guangdong-Dongguan Joint Fund for Basic and Applied Research of Guangdong Province, No. 2020A1515110162, Jianlan Guo, and University Special Fund of Guangdong Provincial Department of Education, No. 2022ZDZX1073, Jianlan Guo.

Data Availability Data availability is not applicable. If needed, the IC surface defect dataset can be obtained by contacting the corresponding author.

Declaration

Conflict of Interest The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

References

1. De Vitis, G.A., Foglia, P., Prete, C.A.: Row-level algorithm to improve real-time performance of glass tube defect detection in the production phase. IET Image Proc. **14**(12), 2911–2921 (2020)
2. Ng, H.-F.: Automatic thresholding for defect detection. Pattern Recogn. Lett. **27**(14), 1644–1649 (2006)
3. Ren, Z., Fang, F., Yan, N., Wu, Y.: State of the art in defect detection based on machine vision. Int. J. Precis. Eng. Manuf.-Green Technol **9**(2), 661–691 (2022)
4. Ciregan, D., Meier, U., Schmidhuber, J.: Multi-column deep neural networks for image classification. In: 2012 IEEE conference on computer vision and pattern recognition. pp. 3642–3649. IEEE (2012)
5. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł., Polosukhin, I.: Attention is all you need. In: Advances in neural information processing systems, vol. 30 (2017)
6. Moutik, O., Sekkat, H., Tigani, S., Chehri, A., Saadane, R., Tchakoucht, T.A., Paul, A.: Convolutional neural networks or vision transformers: Who will win the race for action recognitions in visual data? Sensors **23**(2), 734 (2023)
7. Khan, S., Naseer, M., Hayat, M., Zamir, S.W., Khan, F.S., Shah, M.: Transformers in vision: a survey. ACM Comput Surv (CSUR) **54**(10s), 1–41 (2022)
8. Luan, C., Cui, R., Sun, L., Lin, Z.: A Siamese network utilizing image structural differences for cross-category defect detection. In: 2020 IEEE International Conference on Image Processing (ICIP). pp. 778–782. IEEE (2020)
9. Bennett, M.H., Tobin Jr, K.W., Gleason, S.S.: Automatic defect classification: status and industry trends. In: Integrated Circuit Metrology, Inspection, and Process Control IX, vol. 2439. pp. 210–220. SPIE (1995)
10. Liu, H., Zhou, W., Kuang, Q., Cao, L., Gao, B.: Defect detection of ic wafer based on two-dimension wavelet transform. Microelectron. J. **41**(2–3), 171–177 (2010)
11. Lu, X., Liao, G., Zha, Z., Xia, Q., Shi, T.: A novel approach for flip chip solder joint inspection based on pulsed phase thermography. NDT E Int. **44**(6), 484–489 (2011)
12. Kaitwanidvilai, S., Saenthon, A., Kunakorn, A.: Pattern recognition technique for integrated circuit (ic) pins inspection using wavelet transform with chain-code-discrete Fourier transform and signal correlation. Int. J. Phys. Sci. **7**(9), 1326–1332 (2012)
13. Bergès, C., Soufflet, P., Jadrani, A.: Risk and reliability assessment about a manufacturing issue in a power mosfet for automotive applications. Microelectron. Reliab. **54**(9–10), 1887–1890 (2014)
14. Ding, R., Zhang, C., Zhu, Q., Liu, H.: Unknown defect detection for printed circuit board based on multi-scale deep similarity measure method. J. Eng. **2020**(13), 388–393 (2020)
15. Anvar, A., Cho, Y.I.: Automatic metallic surface defect detection using shuffledefectnet. J. Korea Soc. Comput. Inf. **25**(3), 19–26 (2020)
16. Hu, B., Wang, J.: Detection of PCB surface defects with improved faster-RCNN and feature pyramid network. IEEE Access **8**, 108335–108345 (2020)
17. Hubel, D.H., Wiesel, T.N.: Receptive fields, binocular interaction and functional architecture in the cat's visual cortex. J. Physiol. **160**(1), 106 (1962)
18. Fukushima, K.: Neocognitron: a self-organizing neural network model for a mechanism of pattern recognition unaffected by shift in position. Biol. Cybern. **36**(4), 193–202 (1980)
19. LeCun, Y., Boser, B., Denker, J.S., Henderson, D., Howard, R.E., Hubbard, W., Jackel, L.D.: Backpropagation applied to handwritten zip code recognition. Neural Comput. **1**(4), 541–551 (1989)
20. LeCun, Y., Bottou, L., Bengio, Y., Haffner, P.: Gradientbased learning applied to document recognition. Proc. IEEE **86**(11), 2278–2324 (1998)
21. Gao, X., Jian, M., Hu, M., Tanniru, M., Li, S.: Faster multidefect detection system in shield tunnel using combination of FCN and faster rcnn. Adv. Struct. Eng. **22**(13), 2907–2921 (2019)
22. Xiao, L., Wu, B., Hu, Y.: Surface defect detection using image pyramid. IEEE Sens. J. **20**(13), 7181–7188 (2020)
23. Al-Tam, R.M., Al-Hejri, A.M., Narangale, S.M., Samee, N.A., Mahmoud, N.F., Al-Masni, M.A., Al-Antari, M.A.: A hybrid workflow of residual convolutional transformer encoder for breast cancer classification using digital x-ray mammograms. Biomedicines **10**(11), 2971 (2022)
24. Chen, J., Chen, X., Chen, S., Liu, Y., Rao, Y., Yang, Y., Wang, H., Wu, D.: Shape-former: bridging CNN and transformer via shapeconv for multimodal image matching. Inf. Fusion **91**, 445–457 (2023)
25. Gao, G., Xu, Z., Li, J., Yang, J., Zeng, T., Qi, G.-J.: Ctcnet: a CNN-transformer cooperation network for face image superresolution. IEEE Trans. Image Process. **32**, 1978–1991 (2023)
26. Devlin, J., Chang, M.-W., Lee, K., Toutanova, K.: Bert:Pre-training of deep bidirectional transformers for language understanding, arXiv preprint arXiv:1810.04805 (2018)
27. Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J.D., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., et al.: Language models are few-shot learners. Adv. Neural. Inf. Process. Syst. **33**, 1877–1901 (2020)
28. Huang, L., Tan, J., Meng, J., Liu, J., Yuan, J.: Hotnet: Non-autoregressive transformer for 3D hand-object pose estimation. In: Proceedings of the 28th ACM International Conference on Multimedia, pp. 3136–3145 (2020)

29. Lin, K., Wang, L., Liu, Z.: End-to-end human pose and mesh reconstruction with transformers. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 1954–1963 (2021)
30. Esser, P., Rombach, R., Ommer, B.: Taming transformers for high-resolution image synthesis. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 12873–12883 (2021)
31. Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., et al.: An image is worth 16x16 words: Transformers for image recognition at scale. arXiv preprint arXiv:2010.11929 (2020)
32. Han, K., Xiao, A., Wu, E., Guo, J., Xu, C., Wang, Y.: Transformer in transformer. Adv. Neural. Inf. Process. Syst. **34**, 15908–15919 (2021)
33. Liu, Z., Lin, Y., Cao, Y., Hu, H., Wei, Y., Zhang, Z., Lin, S., Guo, B.: Swin transformer: Hierarchical vision transformer using shifted windows. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 10012–10022 (2021)
34. Chen, C.-F., Panda, R., Fan, Q.: Regionvit: Regionalto-local attention for vision transformers. arXiv preprint arXiv:2106.02689 (2021)
35. Hütten, N., Meyes, R., Meisen, T.: Vision transformer in industrial visual inspection. Appl. Sci. **12**(23), 11981 (2022)
36. Shang, H., Sun, C., Liu, J., Chen, X., Yan, R.: Defectaware transformer network for intelligent visual surface defect detection. Adv. Eng. Inform. **55**, 101882 (2023)

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