# Steel Surface Anomaly Detection Using 3D Depth and 2D RGB Features

Zheng Wangguandong[1], Lu Ping[2], Deng Fangwei[2], Huang Shijun[2], Xia Siyu[1]
(1. Southeast University, Nanjing 210096, China;
2. ZTE Corporation, Shenzhen 518057, China)

DOI: 10.12142/ZTECOM.202601011
https://kns.cnki.net/kcms/detail/34.1294.TN.20260302.1536.002.html, published online March 02, 2026
Manuscript received: 2024-09-27

**Abstract:** The detection of steel surface anomalies has become an industrial challenge due to variations in production equipment, processes, and steel characteristics. To alleviate the problem, this paper proposes a detection and localization method combining 3D depth and 2D RGB features. The framework comprises three stages: defect classification, defect location, and warpage judgment. The first stage uses a data-efficient image Transformer model, the second stage utilizes reverse knowledge distillation, and the third stage performs feature fusion using 3D depth and 2D RGB features. Experimental results show that the proposed algorithm achieves relatively high accuracy and feasibility, and can be effectively used in industrial scenarios.

**Keywords:** anomaly detection; anomaly localization; feature fusion; reverse distillation

**Citation** (Format 1): Zheng W G D, Lu P, Deng F W, et al. Steel surface anomaly detection using 3D depth and 2D RGB features [J]. ZTE Communications, 2026, 24(1): 81 - 87. DOI: 10.12142/ZTECOM.202601011
**Citation** (Format 2): W. G. D. Zheng, P. Lu, F. W. Deng, et al., "Steel surface anomaly detection using 3D depth and 2D RGB features," *ZTE Communications*, vol. 24, no. 1, pp. 81 - 87, Mar. 2026. doi: 10.12142/ZTECOM.202601011.

## 1 Introduction

Steel plates are widely used in various industrial applications. The surface quality of metal products is an important evaluation metric in the metal manufacturing industry[1]. However, metal plates are affected by various factors during manufacturing, such as equipment, processes, and material characteristics. Consequently, surface defects with irregular shapes often emerge[2]. With the development of computer vision, machine vision-based algorithms for detecting such defects have become a research focus[3]. To address challenges in metal surface defect detection—such as scarce defect samples, high variability in shape and type, and the need for precise localization—this paper proposes a steel surface anomaly detection and localization method.

Previous work has introduced various approaches to address challenges in anomaly detection and classification within industrial processes. Zhao et al.[4] proposed a method based on dynamic time warping (DTW) combined with adaptive fuzzy C-means (AFCM), drawing inspiration from similar industrial processes. Wen et al.[5] developed a novel anomaly detection method based on multi-scale knowledge distillation (Ms-KD) and a block domain core information module (BDCI) to quickly screen abnormal images. Yasuno et al.[6] proposed a one-class steel detector using a patch generative adversarial network (GAN) discriminator for visualizing anomalous feature maps.

Fig. 1 shows the pipeline of the proposed method. Given a 2D RGB image, we classify it into two categories (i.e., abnormal and normal) using the data-efficient image Transformer (DEIT) model.

1) For abnormal images, we employ a multi-class DEIT model and a reverse knowledge distillation model to determine the specific defect category and the defect coordinates, respectively.

2) For normal images, we combine the 2D RGB features with the 3D depth map to fuse 2D and 3D features. Furthermore, a new classification head is employed to determine whether the steel plate is warped or flat.

Finally, the defect category is determined by integrating the specific defect type and the warpage status.

## 2 Proposed Method

In actual production, the limited availability of steel defect samples, coupled with the diversity of defect types, poses a significant challenge to accurate detection and classification. Given this data scarcity, it is necessary to achieve efficient discrimination with minimal data. To tackle this, we implement a data-efficient defect multi-classification method for the abnormal multi-classification, which effectively distinguishes be-

This work was supported by ZTE Industry-University-Institute Cooperation Funds under Grant No. HC-CN-20221107001.

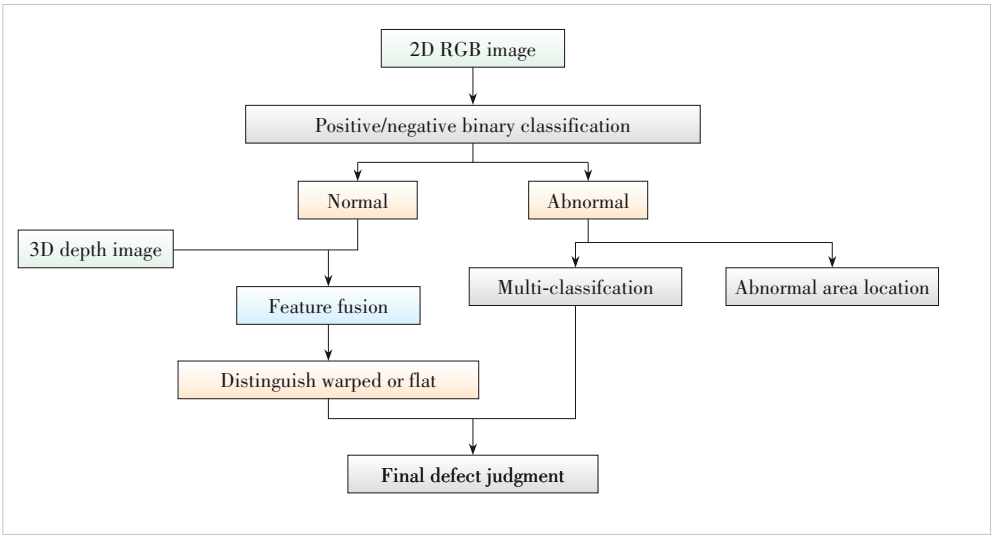


**Figure 1. Pipeline of our proposed method**

tween different types of steel defects. Considering the difficulties traditional knowledge distillation faces in pinpointing defect areas, we employ the reverse distillation defect location method to accurately identify the defective area. Lastly, due to the spatial unevenness of warpage defects, discerning them using only 2D images is challenging. Therefore, we integrate 3D information. Thus, we utilize feature fusion of both 2D RGB and 3D depth images to determine whether the steel is warped.

## 2.1 Data-Efficient Defect Multi-Classification Method

The number of steel defect samples in actual production is limited, necessitating a multi-classification model that can operate effectively with minimal data. Existing transformer-based classification models, such as vision Transformer (ViT), need to be pre-trained on large-scale datasets and then fine-tuned on the ImageNet dataset[7], which requires substantial computing resources. DEIT[8] is essentially a ViT model. It uses three methods: better hyper-parameters, data augmentation and distillation, which can achieve better classification performance with a smaller amount of data.

1) Optimized hyper-parameters. The parameter initialization method chosen is a truncated normal distribution. For learning rate adjustment, we employ a decay strategy: the learning rate first increases linearly during the warm-up phase and then decreases via a cosine method.

2) Data augmentation. A variety of data augmentation methods are used, including random erase, MixUp, CutMix, and exponential moving average (EMA). With MixUp, the resulting images are assigned soft labels rather than single labels. In CutMix, labels are given according to the proportion occupied. EMA ensures that the model weight updates are related to the historical values over time. These methods all help to improve the model's efficiency.

3) Distillation through attention. In the training stage, the class token in ViT for classification is equivalent to an additional patch. It learns the relationship with other patches, and then connects the classifier to calculate CELoss. As shown in Fig. 2, for distillation in DEIT, an additional distill token is added. This token also learns the relationship with other tokens, and then connects the teacher model to calculate KLDivLoss. Subsequently, CELoss and KLDivLoss are combined to form a new loss, which guides the student model training (note that the teacher model is not trained during knowledge distillation).

In the prediction stage, class token and distill token generate different results. These results are then weighted (with a weight of 0.5 each) and summed to obtain the final prediction.

## 2.2 Reverse Distillation Defect Location Method

As shown in Fig. 3, in traditional knowledge distillation, both the teacher and student networks serve as encoders, taking image information as input. The student network learns from the teacher network by reconstructing the representations of the teacher network at different scales[9]. However, in re-

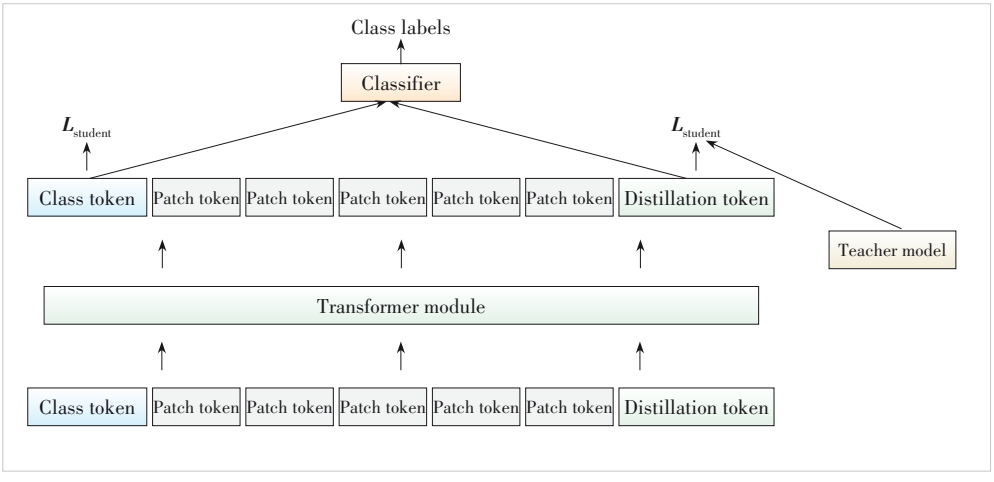


**Figure 2. Distillation token in transformer model**

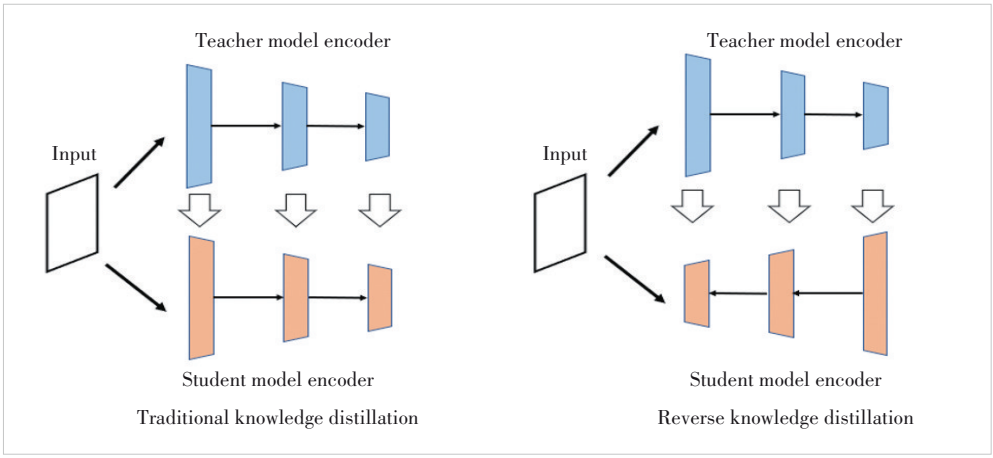


**Figure 3. Difference between traditional knowledge distillation and reverse knowledge distillation**

verse knowledge distillation, the teacher network still acts as an encoder, while the student network functions as a decoder[10]. The low-dimensional features encoded by the teacher model serve as input, allowing the student network to learn the teacher model's representations at different scales by reconstructing them. This process first extracts high-level representations and then refines low-level features. The teacher encoder functions as a downsampling filter, while the student decoder operates as an upsampling filter, creating a symmetric architecture that addresses the limitations of traditional knowledge distillation.

In the inference stage of traditional knowledge distillation, when abnormal samples are input, the student network may reconstruct results highly similar to those of the teacher network. However, to alleviate the problem, Fig. 4 shows that reverse knowledge distillation adopts the following methods:

1) The encoder module (part of the teacher network) utilizes pretrained models. In our implementation, we employed WideResNet and achieved competitive performance.

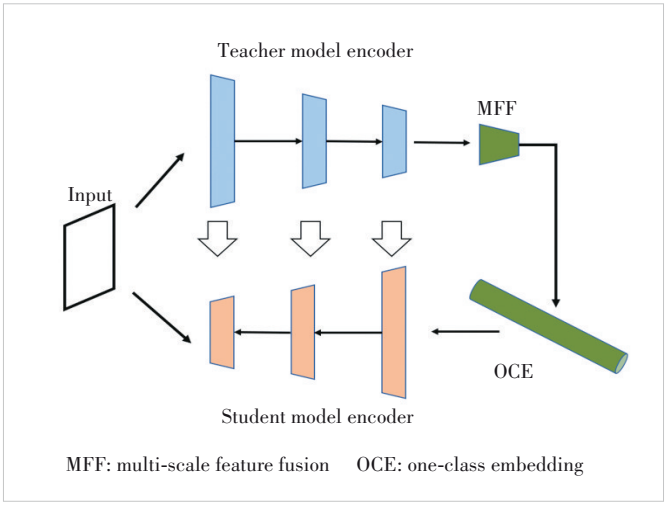


**Figure 4. Reverse knowledge distillation network architecture**

2) A one-class bottleneck embedding (OCBE) module is incorporated into inverse knowledge distillation. It contains a multi-scale feature fusion (MFF) module and one-class embedding (OCE) module. The motivation for employing multi-scale fusion stems from the distinct characteristics of features at different scales: low-dimensional features are rich in texture and edge details, while high-dimensional features encapsulate semantic information. Using only the activation information from the encoder's final layer as the decoder's input could lead to an excess of redundant semantic details. Therefore, by leveraging multi-scale fusion, redundancy is minimized while preserving details.

3) The decoder module mirrors the teacher network but is not an exact copy. During the inference phase, when abnormal samples are input, the reconstructed shapes exhibit greater differences, making the defect features more apparent.

### 2.3 Feature Fusion via 2D RGB Images and 3D Depth Images

With the popularization of various sensors, it is easier to obtain multimodal data from different sources, making it increasingly important to use multi-modal information for various classification and regression tasks[11]. According to how multi-modal fusion is performed, it can be divided into the two following types: aggregation-based fusion and alignment-based fusion.

Aggregation-based fusion employs separate sub-networks to process each modality. The outputs are then aggregated into a unified common feature. These features are subsequently mapped to the output dimension to obtain the final result. The specific formula is as follows:

$$\hat{y}^{(i)} = f\left(x^{(i)}\right) = h\left(\mathrm{Agg}\left(f_1\left(x_1^{(i)}\right),\cdots,f_M\left(x_M^{(i)}\right)\right)\right) \quad (1),$$

where $h$ is the global mapping network, $f$ is the feature extractor, $x^{(i)}$ is the input image, and Agg is the aggregation function. There are many ways to implement aggregation functions, such as averaging multiple modal features and concatenating multiple modal features.

The alignment-based fusion method refers to using an alignment loss to align the features of multiple modalities, retaining the outputs of multiple sub-networks for separate prediction, and finally weighting the prediction results of the different modalities[12]. The optimization objective in this case is shown as:

$$\min \frac{1}{N}\sum_{i=1}^{N} L\left(\sum_{m=1}^{M}\alpha_m f_m\left(x_m^{(i)}\right), y^{(i)}\right) + \mathrm{Alig}_{f_{1,M}}\left(x^{(i)}\right), \text{s.t.} \sum_{m=1}^{M}\alpha_m = 1 \quad (2),$$

where $\text{Alig}_{f_{1:M}}$ is a loss that measures the similarity between two distributions, usually using maximum-mean-discrepancy (MMD). The final output $\sum_{m=1}^{M}\alpha_m f_m\left(x_m^{(i)}\right)$ is an ensemble of $f_m$ associated with the decision score $\alpha_m$, which is learned by an additional softmax output to meet the simplex constraint.

This task focuses on the homogeneous multimodal fusion problem, using 2D RGB images and 3D depth maps for feature fusion. The proposed method belongs to the category of aggregation-based fusion. As shown in Fig. 5, two ResNet18 networks[13] are employed as feature extractors. First, the final fully connected layer of the network is removed; then, 3D depth maps and 2D RGB images are input to extract the two corresponding features. These two features are concatenated, and a new binary classification head is customized to perform the warpage detection task.

## 3 Experiment

We evaluated the performance of this algorithm to classify and locate defects on steel surfaces. Experimental results show the algorithm yields favorable classification and localization outcomes on iron, aluminum, and stainless steel surfaces.

### 3.1 Datasets

1) Defect classification dataset. Two-dimensional defects include abrasions, scratches, holes, stripes and flower patterns, totaling five categories. The materials used are aluminum, iron, and stainless steel. Each full-size steel plate image (2 048×2 048 pixels) is divided into 64 small patches (256×256 pixels). As shown in Fig. 6, all images are collected by 2D line array cameras. The aluminum subset contains 452 abnormal and 153 normal samples, the iron subset contains 307 abnormal and 205 normal samples, and the stainless steel subset contains 243 abnormal and 195 normal samples. All samples are divided into training, verification and test sets at a ratio of 6:2:2.

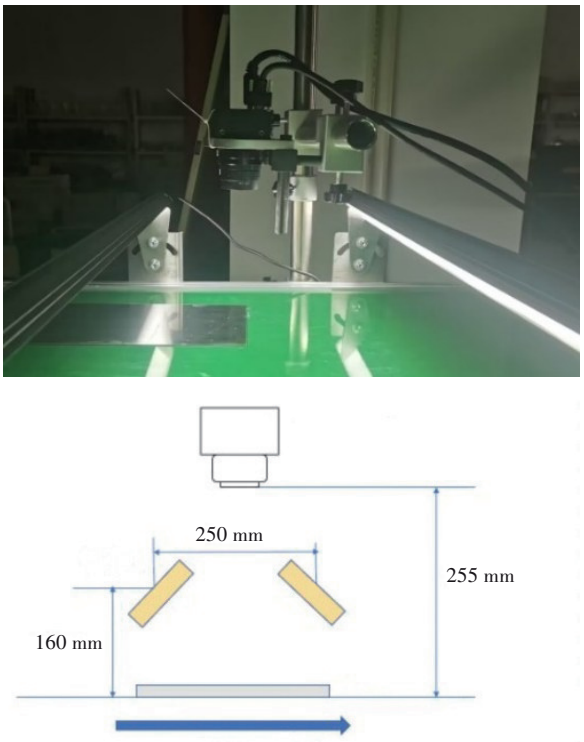


Figure 6. Line array cameras capture real scenes and schematic scenes

2) Defect location dataset. This dataset is constructed similarly to the defect classification dataset, but with the difference that ground truth needs to be incorporated during training to inform the reverse knowledge distillation about the defect locations.

3) Feature fusion dataset. This dataset, comprising 40 2D RGB images and 40 3D depth images, is divided into warping and flattening parts for both modalities. The RGB images are captured of iron plates using a 2D line-scan camera, while the corresponding depth images are acquired using a 3D area-scan camera.

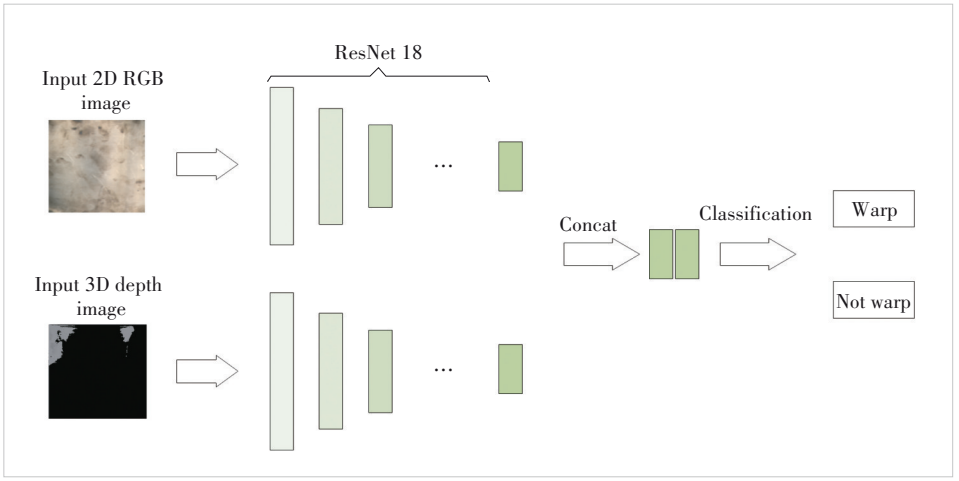


Figure 5. Aggregation-based feature fusion

### 3.2 Implementation Details

In the defect multi-classification process, we employed the DeiT-Tiny pre-training model, using the AdamW[14] optimizer, and the learning rate was set to 5e-5×4/64. The patch size was set to 16 and the LabelSmoothLoss was adopted as the loss function. In the reverse distillation process, WideResNet50[15] was chosen as the teacher model,

trained with the Adam optimizer and a learning rate scheduler (gamma=0.1, with a step size of 5). For feature fusion, ResNet18 was selected as the feature extraction network with a learning rate of 1e-3.

### 3.3 Experimental Results and Analysis

1) Standard of evaluation. Given the characteristics of the classification task, precision (P) is adopted as the primary evaluation metric. True Positives (TP) denote the number of positive samples correctly classified by the model, while False Positives (FP) represent the number of negative samples incorrectly classified as positive.

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}} \quad (3).$$

Based on the characteristics of the location results, the pixel-level Area Under the Receiver Operating Characteristics curve (AUROC) and image-level AUROC are selected as the main location evaluation metrics[16].

AUROC assesses the model's ability to distinguish between positive and negative samples by plotting the false positive rate (FPR) against the true positive rate (TPR) at different thresholds. The FPR represents the proportion of negative samples that are incorrectly classified as positive, while the TPR denotes the proportion of actual positive samples that are correctly identified. Typically, we aim for a high TPR while minimizing the FPR to enhance the model's classification ability.

Pixel-level AUROC evaluates the prediction accuracy of the model at the pixel level, treating each pixel as an independent classification problem. However, the image-level AUROC assesses the entire image for binary classification, disregarding the specific positions and types of each pixel.

2) Data-efficient defect multi-classification. Due to the scarcity of defective steel samples, 25 training sets and 9 verification sets are used to achieve better multi-classification effects on aluminum, iron, and stainless steels. Table 1 shows that the classification accuracy is at least 90% and often reaches 100%. Training loss and validation accuracy are shown in Fig. 7, and the classification results on five types of defects are shown in Fig. 8.

3) Reverse distillation defect location. As shown in Table 2, the pixel-level and image-level AUROC scores indicate high defect localization accuracy. Fig. 7 visualizes these results using heatmaps.

4) Feature fusion via 2D RGB and 3D depth images. An iron plate was selected as the experimental sample. The dataset comprises 20 warped and 20 flat samples for both 2D RGB images and 3D depth images. The data was partitioned into training, verification, and test sets according to a ratio of 6:2:2. Due to the large discrimination of feature representation, the classification accuracy for distinguishing warped from flat samples reached 100%.

**Table 1. Number of testset and precision of aluminum, iron, and stainless steel samples**

| Defect Category | Aluminum | Iron | Stainless Steel |
|---|---|---|---|
| Abrasions | 8/100% | 8/100% | 30/96.7% |
| Scratches | 212/97.17% | 101/99.1% | 18/100% |
| Holes | 28/100% | 12/100% | 8/100% |
| Stripes | 8/100% | 8/100% | 8/100% |
| Flowers | 26/100% | 8/100% | 8/100% |

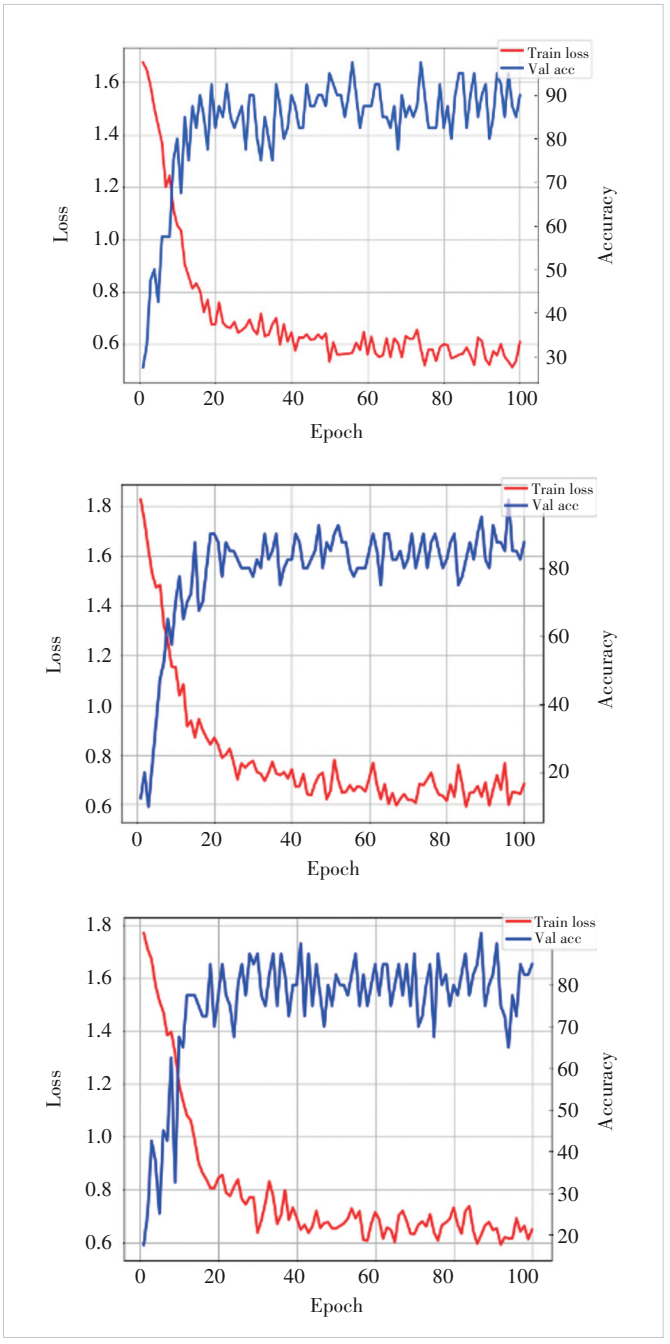


**Figure 7. Training loss and validation accuracy of aluminium, iron, and stainless steel samples, where the red curve means training loss and the blue means validation accuracy**

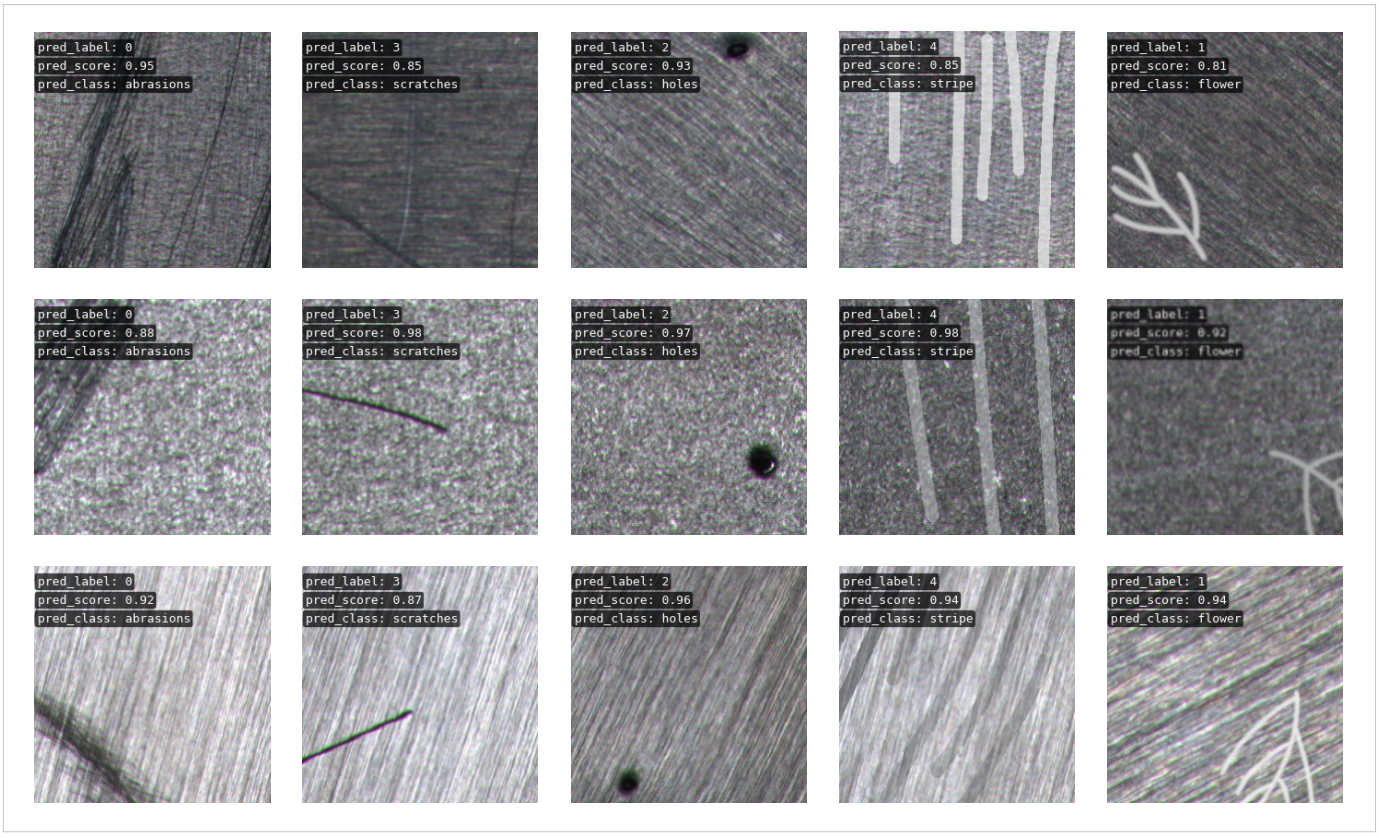


**Figure 8. Classification results of aluminum, iron and stainless steel on five types of defects**

**Table 2. Pixel-level AUROC and image-level AUROC of aluminum, iron, and stainless steel samples**

| Steel Category | Defect Category | Pixel-level AUROC | Image-level AUROC |
|---|---|---|---|
| Aluminum | Abrasions | 97.3% | 96.7% |
| | Holes | 99.8% | 100% |
| | Scratches | 97.5% | 100% |
| Iron | Abrasions | 98.1% | 100% |
| | Holes | 97.8% | 100% |
| | Scratches | 96.3% | 100% |
| Stainless steel | Abrasions | 96.4% | 100% |
| | Holes | 99.2% | 95.4% |
| | Scratches | 85.8% | 98.1% |

AUROC: Area Under the Receiver Operating Characteristics curve

## 4 Conclusions

This paper proposes a method for detecting and locating anomalies of steel surfaces combined with 3D Depth and 2D RGB features, which can be divided into three stages: defect classification, defect location, and warp detection. By leveraging deep learning techniques, the proposed approach minimizes the reliance on manual labor during the inspection process. Experimental results demonstrate that the method achieves the desired accuracy and validates its feasibility.

**References**

[1] Defard T, Setkov A, Loesch A, et al. PaDiM: a patch distribution modeling framework for anomaly detection and localization [C]//ICPR International Workshops and Challenges. ICPR, 2020: 475 – 489. DOI: 10.1007/978-3-030-68799-1_35
[2] Tao X, Gong X, Zhang X, et al. Deep learning for unsupervised anomaly localization in industrial images: a survey [J]. IEEE Transactions on instrumentation and measurement, 2022, 71(1): 1 – 21. DOI: 10.1109/TIM.2022.3196436
[3] He Y, Song K, Meng Q, et al. An end-to-end steel surface defect detection approach via fusing multiple hierarchical features [J]. IEEE transactions on instrumentation and measurement, 2020, 69(4): 1493 – 1504. DOI: 10.1109/TIM.2019.2915404
[4] Zhao J, Liu K, Wang W, et al. Adaptive fuzzy clustering based anomaly data detection in energy system of steel industry [J]. Information sciences, 2014, 259: 335 – 345. DOI: 10.1016/j.ins.2013.05.018
[5] Wen X, Zhao W, Yu Z, et al. A novel anomaly detection method for strip steel based on multi-scale knowledge distillation and feature information banks network [J]. Coatings, 2023, 13(7): 1171. DOI: 10.3390/coatings13071171
[6] Yasuno T, Fujii J, Fukami S. One-class steel detector using patch GAN discriminator for visualising anomalous feature map [PP/OL]. arXiv (2021-06-30) [2025-08-12]. https://arxiv.org/abs/2107.00143
[7] Dosovitskiy A, Beyer L, Kolesnikov A, et al. An image is worth 16×16 words: transformers for image recognition at scale [PP/OL]. arXiv (2021-07-03) [2025-08-12]. https://arxiv.org/abs/2010.11929
[8] Touvron H, Cord M, Douze M, et al. Training data-efficient image transformers & distillation through attention [PP/OL]. arXiv [2025-08-12]. https://arxiv.org/abs/2012.12877
[9] Salehi M, Sadjadi N, Baselizadeh S, et al. Multiresolution knowledge distil-

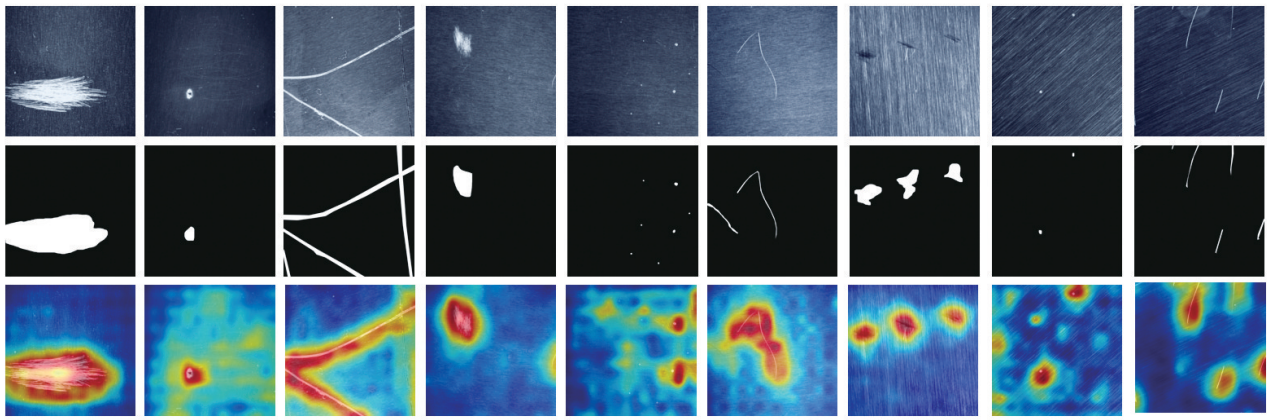

**Figure 9. Defect location results by heatmaps**

lation for anomaly detection [C]//Conference on Computer Vision and Pattern Recognition (CVPR). IEEE, 2021: 14897 – 14907. DOI: 10.1109/cvpr46437.2021.01466
[10] Deng H Q, Li X Y. Anomaly detection via reverse distillation from one-class embedding [C]//Conference on Computer Vision and Pattern Recognition (CVPR). IEEE, 2022: 9727 – 9736. DOI: 10.1109/cvpr52688.2022.00951
[11] Zhu J G, Tang S X, Chen D P, et al. Complementary relation contrastive distillation [C]//Conference on Computer Vision and Pattern Recognition (CVPR). IEEE, 2021: 9256 – 9265. DOI: 10.1109/cvpr46437.2021.00914
[12] Wang Y K, Huang W B, Sun F C, et al. Deep multimodal fusion by channel exchanging [PP/OL]. arXiv [2025-08-12]. https://arxiv. org/abs/2011.05005
[13] He K, Zhang X, Ren S, et al. Deep residual learning for image recognition [C]//Conference on Computer Vision and Pattern Recognition (CVPR). IEEE, 2016: 770 – 778. DOI: 10.1109/CVPR.2016.90
[14] Zagoruyko S, Komodakis N. Wide residual networks [PP/OL]. arXiv [2025-08-12]. https://arxiv.org/abs/1605.07146
[15] Roth K, Pemula L, Zepeda J, et al. Towards total recall in industrial anomaly detection [C]//Conference on Computer Vision and Pattern Recognition (CVPR). IEEE, 2022: 14298 – 14308. DOI: 10.1109/CVPR52688.2022.01392
[16] Kingma D P, Ba J. Adam: a method for stochastic optimization [PP/OL]. arXiv (2014-12-22) [2025-08-12]. https://arxiv.org/abs/1412.6980

## Biographies

**Zheng Wangguandong** is pursuing his master's degree at the School of Automation, Southeast University, China. His research interests focus on artificial intelligence and computer vision, with a specific specialization in image and video generation. He has published four CCF-A conference papers and possesses extensive research experience in image segmentation and object detection.

**Lu Ping** is the executive deputy director of the National Key Laboratory of Mobile Networks and Mobile Multimedia Technology, China. His research areas encompass cloud computing, big data, augmented reality, and multimedia serviceization. He leads and participates in major national science and technology projects and national science and technology support programs. He has published numerous academic papers and is the author of the books *Internet of Things Capability Development and Application* and *Big Data Technology and Application in Cloud Computing*.

**Deng Fangwei** is a senior strategic planner at ZTE Corporation, specializing in industry-specific digital infrastructure, mobile robots, and supporting products for industrial digital transformation.

**Huang Shijun** is a senior strategic planner at ZTE Corporation, with research interests encompassing machine vision, artificial intelligence, computer vision, and deep learning.

**Xia Siyu** (xsy@seu.edu.cn) received his BE and MS degrees in automation engineering from Nanjing University of Aeronautics and Astronautics, China in 2000 and 2003, respectively, and the PhD degree in pattern recognition and intelligence systems from Southeast University, China in 2006. He is currently an associate professor with the School of Automation, Southeast University. His research interests include object detection, applied machine learning, social media analysis, and intelligent vision systems. He was a recipient of the Science Research Famous Achievement Award from the Higher Institution of China in 2015. He has served as a reviewer for many journals including *IEEE T-PAMI*, *T-IP*, *T-SMCB*, *T-IFS*, *T-MM*, and *Neurocomputing*. He received the Outstanding Reviewer Award for Neurocomputing in 2016. He has also served on the PC/SPC for conferences including CVPR, AAAI, ACM MM, and IJCAI. He is a member of the ACM and IEEE.