algorithms

MDPI

Article

Research on a Classification Method for Strip Steel Surface Defects Based on Knowledge Distillation and a Self-Adaptive Residual Shrinkage Network

Xinbo Huang [1,2,*], Zhiwei Song [1], Chao Ji [1], Ye Zhang [1] and Luya Yang [2]

[1] Electronic Information School, Xi'an Polytechnic University, Xi'an 710048, China; 210421136@stu.xpu.edu.cn (Z.S.); jichao@xpu.edu.cn (C.J.); 41703070113@stu.xpu.edu.cn (Y.Z.)
[2] College of Mechanical and Electrical Engineering, Xidian University, Xi'an 710071, China
* Correspondence: huangxb1975@163.com

Abstract: Different types of surface defects will occur during the production of strip steel. To ensure production quality, it is essential to classify these defects. Our research indicates that two main problems exist in the existing strip steel surface defect classification methods: (1) they cannot solve the problem of unbalanced data using few-shot in reality, (2) they cannot meet the requirement of online real-time classification. To solve the aforementioned problems, a relational knowledge distillation self-adaptive residual shrinkage network (RKD-SARSN) is presented in this work. First, the data enhancement strategy of Cycle GAN defective sample migration is designed. Second, the self-adaptive residual shrinkage network (SARSN) is intended as the backbone network for feature extraction. An adaptive loss function based on accuracy and geometric mean (Gmean) is proposed to solve the problem of unbalanced samples. Finally, a relational knowledge distillation model (RKD) is proposed, and the functions of GUI operation interface encapsulation are designed by combining image processing technology. SARSN is used as a teacher model, its generalization performance is transferred to the lightweight network ResNet34, and it is conveniently deployed as a student model. The results show that the proposed method can improve the deployment efficiency of the model and ensure the real-time performance of the classification algorithms. It is superior to other mainstream algorithms for fine-grained images with unbalanced data classification.

Keywords: classification of strip steel defects; adaptive residual shrinkage network; relational knowledge distillation; Cycle GAN data enhancement; unbalanced data; image processing

Citation: Huang, X.; Song, Z.; Ji, C.; Zhang, Y.; Yang, L. Research on a Classification Method for Strip Steel Surface Defects Based on Knowledge Distillation and a Self-Adaptive Residual Shrinkage Network. *Algorithms* **2023**, *16*, 516. https://doi.org/10.3390/a16110516

Academic Editor: Laura Antonelli

Received: 12 October 2023
Revised: 25 October 2023
Accepted: 8 November 2023
Published: 10 November 2023

1. Introduction

Strip steel is an essential category of steel. Due to the limitation of production technology, many defects such as inclusions, patches, and scratches will appear on the zsurface of strip steel. It is worth noting that different types of defects have different effects on the service life of strip steel and may even cause safety accidents. To ensure the production quality of strip steel, different types of defects must be identified in the production process of strip steel [1]. However, at present, the development of strip surface defect detection technology is slow, and many production lines still adopt manual classification or traditional image classification methods, which cannot cope with problems such as loud noise, low contrast, unbalanced samples, and difficult segmentation of defects [2].

Currently, three kinds of image classification algorithms based on computer technology exist. One is the classification method based on image processing technology. Chagas et al. [3] proposed an image texture analysis method integrating the probability distribution based on image processing and information theory technology. The features extracted by this method are input into the classifier with remarkable effect, but this approach is not effective for processing large-scale data. Second, a traditional machine learning algorithm is used for feature extraction and classification. Chu et al. [4] adopted an anti-noise support

vector hypersphere classifier to solve the classification problem of steel surface defects, and the classification accuracy reached 96.88%. Xu et al. [5] had several SVM models built by AdaBoost and sonar images with low resolution and noise were classified for the first time. However, the feature distribution of some surface defects is irregular and difficult to extract. Traditional machine learning methods easily fall into a local optimum based on predetermined rules, and it is difficult to improve the classification accuracy. Another approach is to extract the surface features of objects by convolutional neural networks. Ju et al. [6] proposed a method called synergic adversarial label learning (SALL) which used the knowledge distillation cooperative training model to realize the task of target classification. Guo et al. [7] combined channel and spatial attention mechanisms for designing lightweight attention mechanism modules that can be easily embedded into any convolutional neural network (CNN) to improve the fine-tuning efficiency of the network. However, there are still two problems to be solved. First, the above methods require many data resources and cannot overcome the problem of few defective samples. Second, while ensuring the classification accuracy, the methods cannot meet the real-time requirement of model deployment and detection in the industrial production field. In this regard, Chiu et al. [8] integrated the Mask-RCNN framework and copy-paste data enhancement method to accomplish mixed type defect classification under the condition of insufficient defect samples, and the accuracy of the single model reached 97.7%. Tu et al. [9] improved the real-time performance of the detection network through the TensorRT framework, and its classification accuracy reached 93.5%. Although data enhancement technology is widely used in few-shot learning [10], traditional data enhancement methods such as random rotation, flipping, image translation, and random crossing can enhance the dataset to a certain extent, but the diversity of the dataset is difficult to change. For this purpose, Lv et al. [11] migrated data from existing defect samples by generating adversarial networks, but this method is not effective in enhancing the local information of images, and it is highly dependent on the datasets. Obviously, none of the methods proposed in [8–11] can resist the imbalance of data samples. In this regard, Lerner et al. [12] combined the naive Bayes classifier and multilayer perceptron neural network to solve the multiclassification problem of unbalanced samples in high-dimensional mode, and the effect was remarkable. On this basis, Yuan et al. [13] proposed a learning method called uncorrelated cost sensitive multiset learning (UCML) to solve the problem of highly unbalanced data classification. This method enhances the robustness of the model by constructing multi-set learning discriminant features and introducing generating adversarial networks to fit the original data distribution.

However, relying only on deep learning will lead to overreliance on labeled samples, an approach which cannot be universally applied to all data. Therefore, in recent years the application of multisource information fusion technology in the field of image classification has become more popular. Liang et al. [14] proposed the attention multisource fusion method. In the field of meta-learning, the problem of few-shot learning is solved by means of class alignment, domain attention distribution, and multisource data fusion. Li et al. [15] used the multiscale information feature fusion method to re-extract the texture information of the bottom layer of the image to focus more attention upon the slender and easily overlooked defects of the strip steel surface, and the detection accuracy of this method reached 98.26%. Wang et al. [16] fused the historical data of strip defects to track model faults, which can assist expert decision-making.

After comparison, the method proposed in [4,5] cannot fully extract the fine-grained information of the image, thus ignoring the local details. Although the feature extraction ability of the model is improved through transfer learning and the attention mechanism, it relies too much on the distribution fitting of the original data [6,7]. The works of [8–11] overcome the problem of insufficient defective samples through data enhancement technology, but it cannot cope with the unbalanced data distribution. The method of [13] can combat uneven samples, but it ignores the feature extraction ability of backbone networks and the efficiency and deployment of the models. Therefore, inspired by references [14–16], this pa-

per integrates fine-grained information extraction, data sample enhancement, unbalanced sample confrontation, and model generalization performance transfer.

To solve the problems of fewer samples and unbalanced data in industrial strip production and meet the requirements of real-time classification on site, a self-adaptive residual shrinkage network classification method based on relational knowledge distillation was proposed. The main contributions are summarized as follows:

- In the aspect of data enhancement, aiming at the shortage of defect samples in a single production line, a data enhancement method of image mosaic fusion is proposed to improve the matching degree between the model and the data. Aiming at the problem of insufficient defect samples across production lines, an image generation method based on Cycle GAN is proposed to realize the cross domain conversion of defect samples. Compared with the limitations of [8–11], this method solves the problem of insufficient strip steel defect samples from two angles;
- In terms of feature extraction, based on the integration of the attention mechanism [7], a new backbone network: the self-adaptive residual shrinkage network (SARSN) is proposed to solve the difficulty of image fine-grained feature extraction through a soft threshold and the channel attention mechanism;
- For unbalanced samples, a new adaptive loss function is designed to manage the sample categories separately to achieve a balance between classes to improve the accuracy of the classification model. Compared with the existing methods [12,13], this method is better at handling interclass differences and reduces the sensitivity of the model to data;
- In terms of model deployment efficiency, compared with the literature [6], this method focuses on optimizing the network structure through knowledge distillation and transferring the generalization performance of a large-scale network model to a small-scale lightweight network While ensuring the real-time performance of the classification network, the classification accuracy is 4.3% higher than that in the literature [9];
- Finally, this paper quantifies the evaluation index of strip steel defects by image processing technology and designs a GUI interface that is convenient for users to operate.
- The organization of the other sections of this paper is as follows: the second part is the theoretical description of the relevant methods. Section 2.1 presents the deployment of the whole algorithm and the algorithm design process. Section 2.2 describes the principle of Cycle GAN data enhancement. Section 2.3 introduces the design process of each part of the feature extraction backbone network. Section 2.4 puts forward the theoretical basis of structured relational knowledge distillation. The third part is the description of the experimental process. Section 3.1 introduces the image preprocessing process; Section 3.2 verifies the performances of the teacher model and student model, respectively. Comparative experiments are carried out in Sections 3.2.3 and 3.2.4. Finally, combined with an image processing algorithm, the defect evaluation index is proposed, and the GUI operation interface is designed. The fourth part presents the conclusion and prospects.

2. The Proposed Theory

2.1. Model Deployment Process and Algorithm Structure Design

In the production process of strip steel, the applied classification algorithms should not only ensure its accuracy, but also meet the real-time requirements of the industrial field, which imposes high requirements for the model deployment process and the design of the classification algorithms. Figure 1a shows the deployment flowchart of the classification model for strip steel defects. In the preprocessing stage, first, the strip steel defect images are obtained. Through the processor, they are input into the classification network, and the defects are classified by related processing algorithms. The detection speed of the classification algorithms matches the running speed of the strip conveyor. Finally, the classification results are input into the human–computer interaction system for users to operate. As shown in Figure 1b, the strip defect classification algorithm is designed for

this paper, and the strip steel defect image is first preprocessed. To ensure sufficient data, image mosaicing and fusion and data enhancement based on Cycle GAN are carried out. Second, a deep network model is designed: a self-adaptive residual shrinkage network (SARSN) and the self-adaptive directional derivative threshold (SADDT) module constitute the core of the network. Then, a dropout layer, a dense layer and a fully connected layer are added. Next, the results are input into the softmax classifier. The network is used as the teacher model in relational knowledge distillation, and the student model is deployed in the front end with the lightweight network ResNet34 to learn the generalization ability of the teacher model. Finally, the student model is trained and tested.

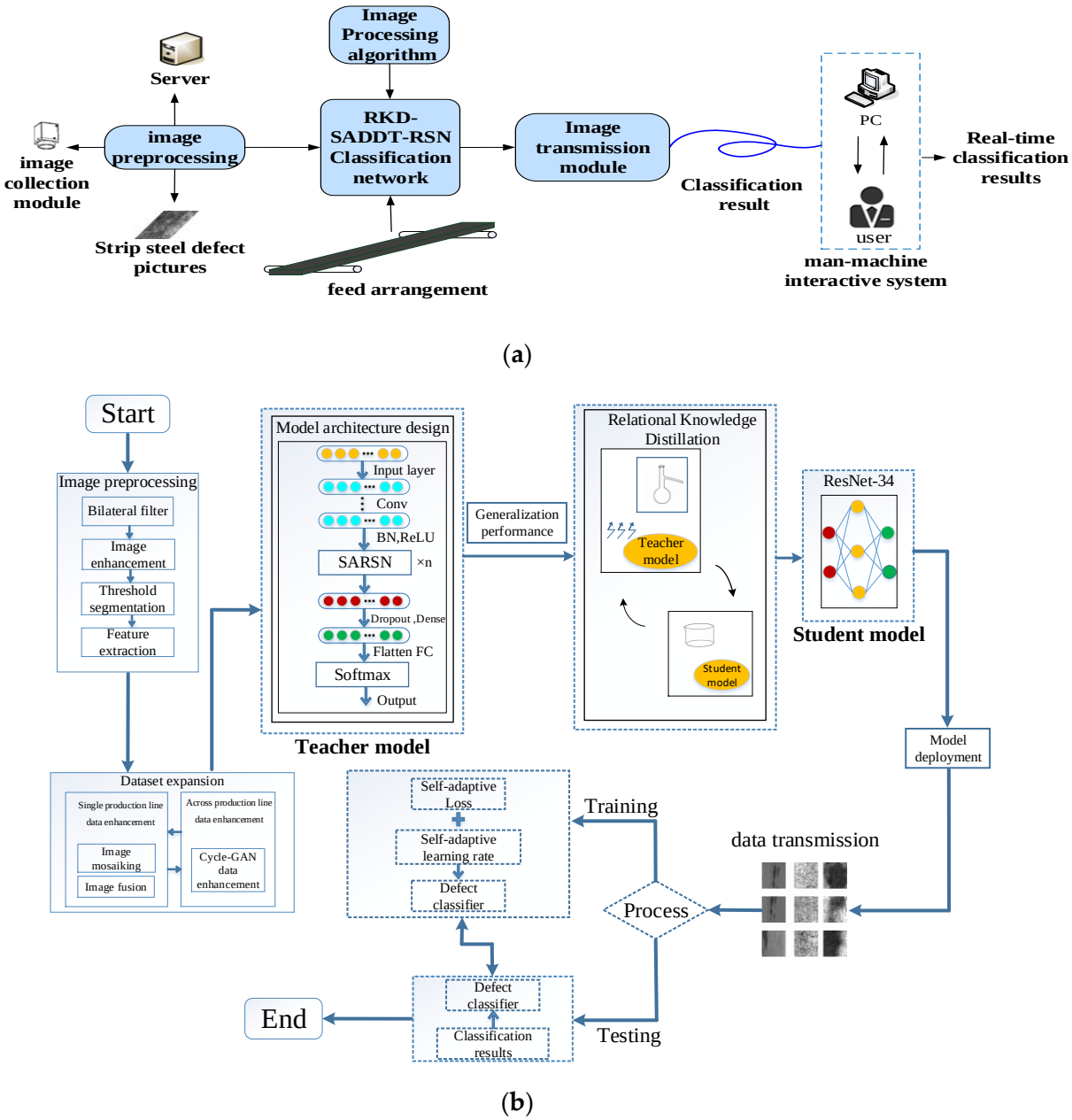


Figure 1. (**a**) Model deployment process. (**b**) The framework of the proposed method.

2.2. Image Cross-Domain Conversion: Data Enhancement Based on Cycle GAN

Due to the significant differences of images in different production lines, it is difficult to directly apply the sample defects collected by the existing production lines to other production lines, resulting in the poor universality of defect samples and difficulties in

deploying them in other production lines. To solve the above problems, this paper proposes a sample migration method based on Cycle GAN for generating adversarial networks. The existing defective samples are migrated, applied to the new production line, and combined with non-defective samples. Many surface defect sample data suitable for the new production line are obtained.

The basis of defect sample data migration comes from the similarity of the underlying features of different images, including lines, colors, textures, and other features. The CNN can thoroughly learn these image underlying features in source domain A and migrate them to target domain B. A schematic diagram of defect sample migration algorithms is shown in Figure 2.

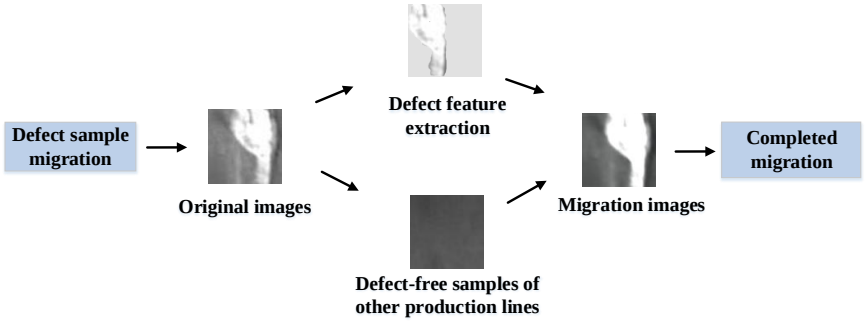


Figure 2. Defect sample migration algorithms.

Figure 2 connects the defect samples obtained from production line 1 with the defect-free samples from production line 2 to obtain new defect samples.

The generative antagonistic network was proposed by Ian Goodfellow in 2014 [17] through confrontation training between a generative model (generator) and a discriminator model (discriminator) to generate images that are false and genuine. The generator G and the discriminator D are continuously optimized to obtain the optimal solution. In this process, the authenticity of false samples generated by the generated model increases, and the discrimination ability of the discriminant model becomes stronger. Unlike the classical GAN algorithms, Cycle GAN has an obvious effect in solving the problem of unlabeled and unpaired image cross-domain conversion, and its model includes double mapping G: $X \rightarrow Y$ and F: $Y \rightarrow X$. The loss function of Cycle GAN is the adversarial loss function [18], which is defined based on generator function G and discriminator D_Y.

$$L_{GAN}(G, D_Y, X, Y) = E_{y \sim P_{data}(y)}[logD_Y(y)] + E_{x \sim P_{data}(x)}[log(1 - D_Y(g(x))]. \quad (1)$$

where D_Y is used to distinguish whether the data come from the generated sample $G(x)$ or the true sample y, E is the expectation, G is used to minimize the adversarial loss function $L_{GAN}(\ )$, and the function of D is opposite to that of G. The goal of the adversarial loss function is to learn the interdomain mapping of $X \rightarrow Y$, where $\{x_i\}_{i=1}^{N}, x_i \in X$, $\{y_i\}_{i=1}^{N}, y_i \in Y$. The data distribution is $x \sim p_{data}(x)$, and $y \sim p_{data}(y)$, and $X \rightarrow Y$. The role of the adversarial loss function is to match the distribution of the generated image with the target domain, but the confrontation loss cannot completely map the input to the output y_i', so we define the cycle consistency loss [19] to further reduce the space of the mapping function, and the cycle consistency loss function is:

$$L_{cyc}(G, F) = E_{x \sim p_{data(x)}}[||F(x) - x||_1] + E_{y \sim p_{data(x)}}[||G(y) - y||_1]. \quad (2)$$

The total loss function is:

$$\begin{aligned} L(G, F, D_X, D_Y) &= L_{GAN}(G, D_Y, X, Y) + L_{GAN}(F, D_X, Y, X) \\ &+ \lambda_1 L_{cyc}(G, F) + \lambda_2 L_{id}(G, F). \end{aligned} \quad (3)$$

In Equation (3), λ_1 and λ_2 are used to control the importance of confrontation loss and circulation consistent loss, the purpose is:

$$G^*, F^* = \underset{G,F}{argmin} \underset{D_X D_Y}{max} L(G, F, D_X, D_Y) G^* \quad (4)$$

The structural schematic diagram of Cycle *GAN* is shown in Figure 3.

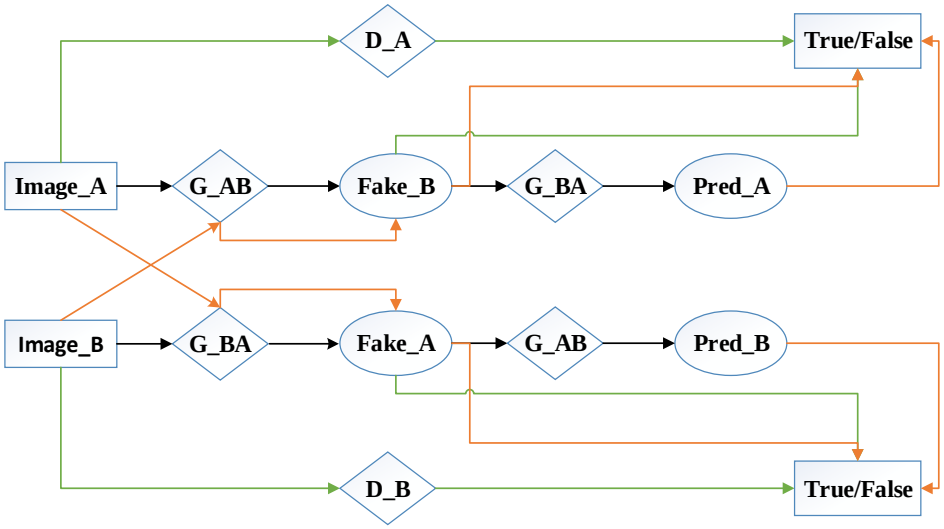


Figure 3. Cycle GAN structure diagram.

There are four network structure routes to be trained in the figure, which are A→A, A→B→A, B→B, and B→A→B. Among them, A→B→A and B→A→B can transfer in two directions, and the whole Cycle GAN network completes the bidirectional transfer of the sample style through cyclic mapping and establishing cycle consistency loss.

2.3. Classification Network Structure Design

2.3.1. Deep Residual Shrinkage Network

The proposed residual learning framework [20] solves the problem of model degradation with the deepening of neural network layers. Shot-cut connections are added between convolution layers to form a basic residual unit (RBU) [21], which is shown in Figure 4a,b, and the overall structure diagram of ResNet is shown in Figure 4c. The structure consists of an input layer, a convolution layer, batch normalization (BN), some RBUs, ReLU, global average pooling (GAP), and a fully connected (FC) layer, which will serve as the basis for further research in this paper. Residual learning improves accuracy and avoids gradient disappearance by increasing the network depth without adding additional parameters and computational complexity. Similar to the critical position of RBUs in ResNet, the residual shrinkage building unit (RSBU) is the core part of the deep residual shrinkage network [22]. As shown in Figure 4d, the RSBU structure consists of two convolution layers, two ReLUs, two BNs, a soft threshold module, and an identity shortcut, in which global average pooling (GAP) obtains a set of one-dimensional vectors and inputs them into the fully connected layer (FC) and then maps the input of the FC layer to the range of (0, 1) through the sigmoid activation function, which maintains good output characteristics.

K represents the number of convolution kernels in the convolution layer, *C* represents the number of channels in the feature map, *W* represents the width of the feature map, and 'C × W × 1' is the product of the number of channels and the width and height of the feature map.

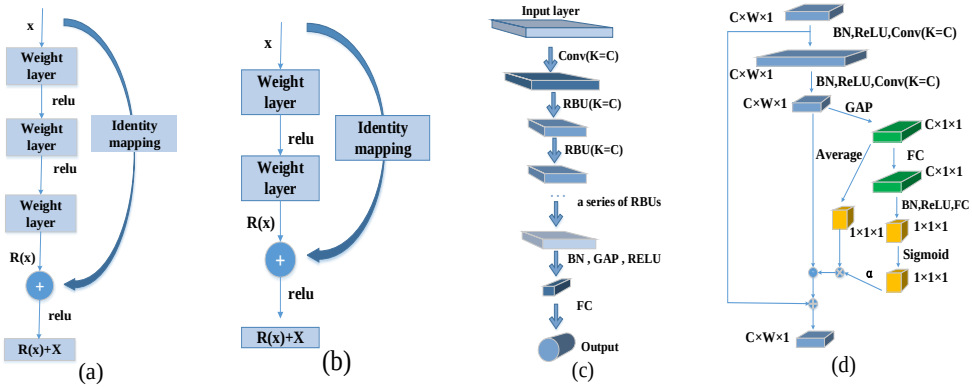


Figure 4. Schematic diagram of the residual learning unit. (**a**) Building block, (**b**) bottleneck, (**c**) architecture of a Res Net, and (**d**) architecture of RSBU.

2.3.2. The Proposed Self-Adaptive Residual Shrinkage Network (SARSN)

Based on RSBU, this paper proposes combining the attention mechanism module and self-adaptive directional derivative threshold module (SADDT) to obtain an adaptive residual shrinkage network (SARSN).

Figure 5 shows that this paper adds an attention mechanism branch after two BNs, two ReLUs, and two convolution layers, which can automatically infer a suitable parameter, namely the self-adaptive directional derivative (SADDT), and the function of the soft threshold function is replaced by the SADDT. Parallel branches can make the neural network direct more attention to details effectively, and finally import into the backbone output of the neural network is ultimately imported through the feed-forward channel. A schematic diagram of the SARSN backbone network is shown in Figure 6.

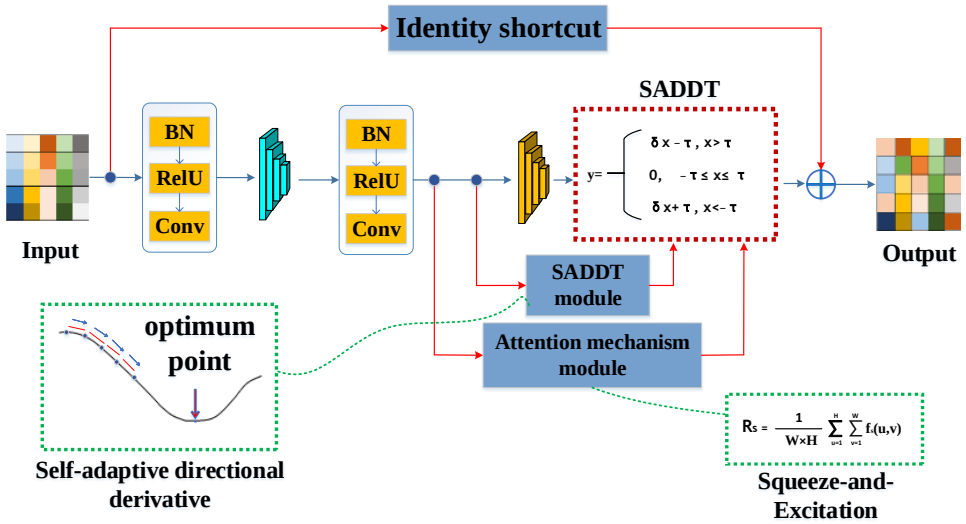


Figure 5. The architecture of SARSN.

The input channel is 1 × 3 × 64 × 64. After entering the SARAN backbone network, the feature map of 1 × 64 × 64 × 64 is first obtained through the Conv, ReLU, and BN layers, and then the feature map of 1 × 512 × 16 × 16 is sequentially obtained through Sequential composed of four different numbers of BasicBlock, and the final the output is determined to be 1 × 512 × 1 × 16 through GAP.

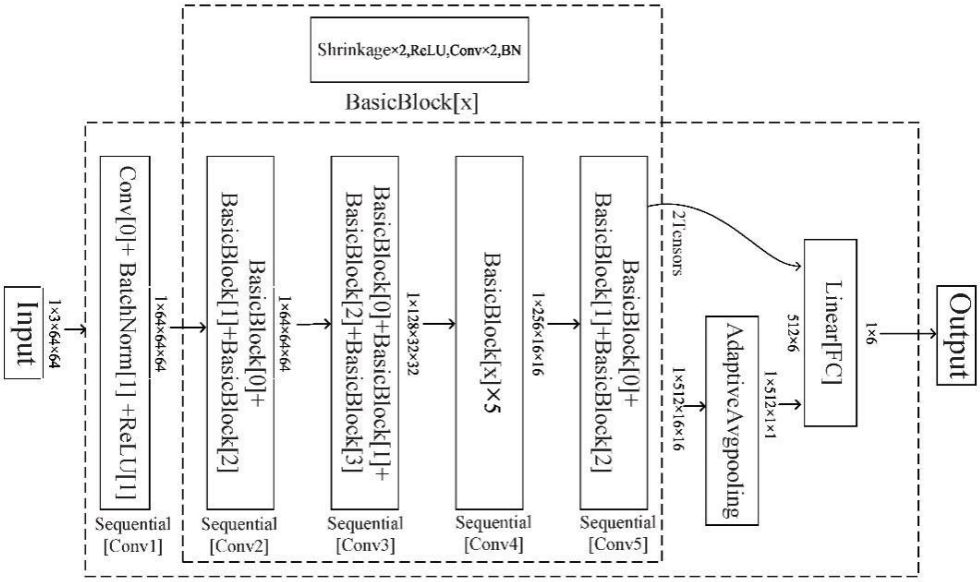


Figure 6. Schematic diagram of the SARSN backbone network.

(1) Attention mechanism and squeeze-and-excitation networks:

At present, attention mechanisms in the field of computer vision mainly include squeeze-and-excitation networks [23], spatial transformer networks (STN) [24] and the convolutional block attention module (CBAM) [25]. The method proposed in this paper uses SE-Net to learn a set of correlation weights to adjust the size of each feature channel, which is equivalent to giving each feature channel different degrees of attention. The network can selectively pay attention to some local information and ignore useless information. The model adopts a small subnetwork structure without introducing extra parameters and computational complexity. The statistic R is a value obtained by shrinking f in the $W \times H$ dimension space. The calculation formula of the parameter R is shown in Formula (5)

$$R_s = \frac{1}{W \times H} \sum_{u=1}^{H} \sum_{v=1}^{W} f_s(u, v). \tag{5}$$

where f and R represent the input and output of the extrusion operation, respectively, u, v, and s represent the width, height and channel number of the feature map, and W and H represent the width and height of the input feature map, respectively. To quickly extract helpful feature information during the extrusion operation, the network not only needs to obtain the combination of nonlinear relationships among channels, but also needs to ensure that multiple channels can be activated. Generally, the bottleneck structure is selected to complete this operation. First, the dimension is reduced from c to c/r, and the output signal is matched with the input. The calculation process is shown in Formula (6)

$$\alpha = \sigma(W_2 ReLU(W_1 t)). \tag{6}$$

where $W_1 \in \mathbb{R}^{c \times \frac{c}{r}}$, $W_2 \in \mathbb{R}^{\frac{c}{r} \times c}$ represent the weight and deviation values, respectively, $\sigma(\ldots)$ and $ReLU(\ldots)$ are activation functions and α is the output of excitation, which reflects the importance of each channel. The output of the SE loop obtained by activating α is shown in formula (7):

$$\widetilde{x}_n = e_n x_n. \tag{7}$$

where $\widetilde{x}_n = [\widetilde{x}_1, \widetilde{x}_2, \ldots, \widetilde{x}_n]$ is the output of the nth feature map and e_k means the weight, which represents the priority of the nth pass. In the SE loop, the excitation maps the parameter t to a set of channel weights, which can be regarded as a channel self-attention

function in essence. The SE-Net structure includes three parts: the squeeze operation, excitation operation, and reweight operation.

(2) Self-adaptive directional derivative threshold

A soft threshold [22] can eliminate noise interference in vibration signals in industrial fault diagnosis [26], and the soft threshold function can decompose input signals and filter signals within threshold τ. This method ignores parts with absolute values less than τ and shrinks parts with absolute values greater than τ. The soft threshold function is shown in Formula (8):

$$y = \begin{cases} x - \tau, & x > \tau \\ 0, & -\tau \le x \le \tau \\ x + \tau, & x < -\tau \end{cases} \tag{8}$$

where x, y and τ represent the input, output, and threshold, respectively. In addition, the gradient values of the soft threshold function output after derivation of the input are 0 and 1. This method can effectively avoid the phenomenon of gradient disappearance and gradient explosion. The partial derivative calculation is shown in Formula (9)

$$\frac{\partial y}{\partial x} = \begin{cases} 1, & x > \tau \\ 0, & -\tau \le x \le \tau \\ 1, & x < -\tau \end{cases} \tag{9}$$

In the field of image classification, different types of noise, such as Gaussian white noise and salt and pepper noise, will also be introduced in the process of industrial scene image acquisition and transmission. When the soft threshold function processes pixel information, it also retains the features outside the threshold interval, in which noise interferes with the image quality. This leads to the reduction of classification accuracy in some fine-grained feature image classification tasks [27]. Therefore, this paper proposes a self-adaptive directional derivative threshold (SADDT) to optimize the deficiency of the soft threshold in fine-grained image classification. SADDT deduces the derivative value of image feature information in the appropriate direction through the channel attention mechanism. Like the optimal gradient in the gradient descent algorithms, SARSN can learn helpful information in the threshold interval instead of keeping all the information in the threshold interval. The SADDT function is shown in Formula (10)

$$y = \begin{cases} \delta x - \tau, & x > \tau \\ 0, & -\tau \le x \le \tau \\ \delta x + \tau, & x < -\tau \end{cases} \tag{10}$$

where δ is the directional derivative value, because the noise situation in different image samples is quite different and an appropriate threshold should be chosen for each sample according to the situation. Therefore, this paper takes the output product of the squeeze operation and excitation operation as a new threshold, and adjusts an appropriate threshold as shown in Formula (11)

$$P_s = \underset{u,v}{Mean}(|sigmoid(\widetilde{x}_{u,v,s})|) \cdot R_s. \tag{11}$$

where $\widetilde{x} \in \mathbb{R}^{W \times H \times C}$ and $R \in \mathbb{R}^{C}$ represent the outputs of the rescale operation and squeeze operation, respectively. $P \in R^{C}$ indicates the threshold corresponding to each channel in the feature map.

(3) Self-adaptive loss function

At present, most loss functions basically do not consider the differences between classes when dealing with multiclassification tasks, which makes the resampling process of imbalanced sample datasets difficult [28] and data with a small sample size will be overfitted. To improve the accuracy of unbalanced samples, all samples can be clustered

into most categories. In this paper, an adaptive update classification loss method based on accuracy and geometric mean (*GMean*) [29] is proposed. This method achieves the effect of interclass balance by treating each category separately, and *GMean* is the geometric average of recall and specificity. Its calculation is shown in Formula (12):

$$GMean = \sqrt{TPR \times FPR}. \tag{12}$$

where *TPR* is the true positive rate and *FPR* is the false positive rate. The expression of the weight factor ϕ for each category of adaptive update loss distribution is shown in Formula (13):

$$\Phi = K \times exp(-\frac{GMean}{2}) \times exp(-\frac{1 - accuracy}{2}). \tag{13}$$

where ϕ is the weight factor of each category, and k is the most significant proportion of a few sample datasets, that is the sample imbalance ratio. Accuracy is the classification accuracy; then, $(1 - accuracy)$ is the error rate, its calculation is shown in Formula (14):

$$accuracy = \frac{TP + TN}{TP + TN + FP + FN}. \tag{14}$$

where *TP* denotes the true positive, which is the number of positive classes with correct classification, and FN denotes the false negative which is the number of negative classes with incorrect classification. *FP* and *FN* can be deduced in the same way.

To avoid the phenomenon of gradient explosion caused by the weight of a few samples being too large in the dataset, this paper uses the tanh activation function to map the value of ϕ between -1 and 1, and its expression is shown in Formula (15):

$$T_h = \frac{exp(1 - \Phi) - exp(\Phi - 1)}{exp(1 - \Phi) + exp(1 + \Phi)}. \tag{15}$$

According to Formula (15), the weight of a few samples is between -1 and 1, while the weight of most samples is close to 0. This operation not only avoids the gradient explosion phenomenon, but also does not reduce the sensitivity of the loss function when dealing with an unbalanced data distribution [30]. The accuracy between categories can be measured by the Euclidean distance between the real sample value and the predicted value [31]. The square of the Euclidean distance is usually used in research, as shown in Formula (16):

$$\rho^2 = \sum_{i=1}^{n} (y_i - y_i' + \beta)^2. \tag{16}$$

where ρ is the square of the Euclidean distance, y_i represents the true value, $y_i{}'$ represents the predicted value, and n represents the sample size of each mini-batch, and β is the parameter to adjust the sensitivity of the loss function. The expression of the loss function is shown in Formula (17):

$$Loss(\theta) = -\frac{\rho^2}{n} \sum_{i=1}^{n} T_h log(y_i{}'). \tag{17}$$

where θ is a parameter that can emphasize a few samples in the training process, increase the classification accuracy of a few samples, and cooperate with the optimizer to realize the global optimization of the neural network.

2.4. Knowledge Distillation

The rapid development in the fields of computer vision and deep learning benefits from today's advanced computing power, which can support the deepening of the number of neural network layers. Usually, the models with better effects have high computing cost requirements. However, in many cases, the convenience of model deployment is highlighted while keeping the model unchanged. The chemical concept of "distillation" can effectively separate the components with different boiling points in complex mixtures.

Based on the concept of model compression, Hinton et al. [32] proposed the concept of knowledge distillation (KD) by training a more complex teacher model (called distillation), allowing the student model to adequately fit the teacher softmax output distribution of the model, thus migrating knowledge from complex models to smaller models that are more easily deployed.

At the same time, to meet the requirements of fine-grained feature extraction accuracy and real-time model deployment in the classification of strip steel defects, this paper proposes the method of relational knowledge distillation (RKD) [33], which changes the point-to-point output mode of the above traditional knowledge distillation methods and completes knowledge transfer through the conversion relationship between data instance structures. The traditional distillation expression is shown in Formula (18):

$$F_{T,S}(q_i) = \left(\frac{exp(\frac{F_T(z_i)}{\tau})}{\sum_j exp(\frac{F_T(z_j)}{\tau})}, \frac{exp(\frac{F_S(z_i)}{\tau})}{\sum_j exp(\frac{F_S(z_j)}{\tau})}\right). \tag{18}$$

where q_i is the distribution generated by the model output, z_i is used to generate the class probabilities, and τ is the temperature, which is similar to the Boltzmann distribution in statistical mechanics. The geometric mean of the class probability distribution generated by z_{ij} in the teacher model will be used as the "soft target" to train the student model. When the "soft target" has higher entropy, the gradient variance between data distributions will be reduced. The goal of model training is to minimize the cross-entropy between data distributions. Relational knowledge distillation (RKD) completes the knowledge transfer through the correspondence structure between the data, and the process of knowledge transfer is shown in Figure 7.

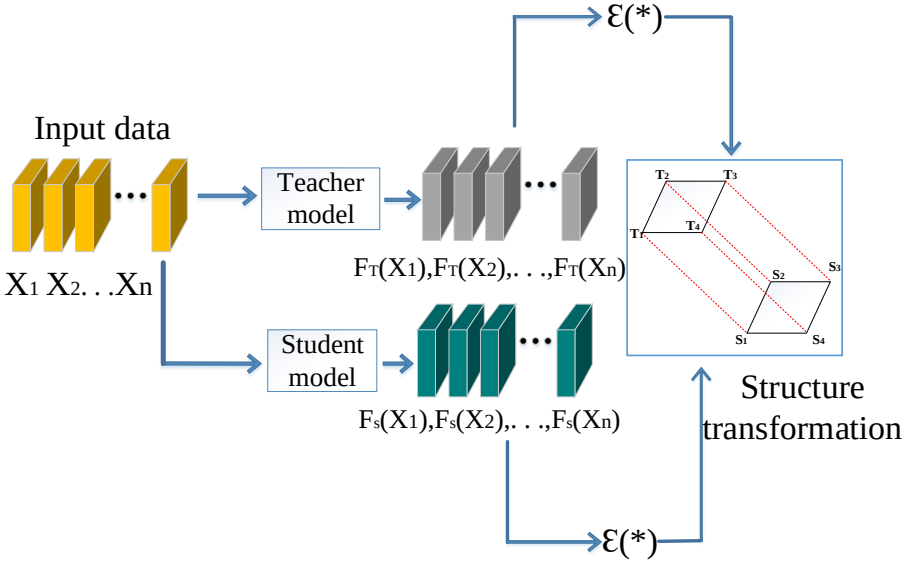


Figure 7. Relational knowledge distillation.

It can be seen from Figure 7 that unlike conventional knowledge distillation, the loss function of RKD does not directly act on the output of the teacher model but cascades the relationship information of N-ary arrays through the potential energy function $\tau-$s (the gray and green parts in the figure represent two groups of cascaded model functions). $\tau-$s is the bridge connecting the teacher network and the student network, and after the input Xn passes through the loss function, by mapping information, the dimensional differences caused by knowledge transformation can be eliminated. Therefore, the information transfer process described in the figure is parallel rather than serial. The introduction of $Loss_{RKD}$ will be carried out below.

To avoid single data output, the student cluster Sa and the teacher cluster Ta form a structural relationship, which is more conducive for the student model to learn the representation information of the image and embed it into the corresponding spatial structure. The expression of the KD loss function is shown in Formula (19), and its basic form comes from Formula (20):

$$Loss_{KD} = \sum_{x_i \in \lambda} l(F_T(x_i), F_S(x_i)). \tag{19}$$

$$Loss_{RKD} = \sum_{(x_1, x_2, \ldots, x_n) \hat{I} \lambda^N} l(\varepsilon(F_T(x_1), \ldots, F_T(x_n)), \varepsilon(F_S(x_1), \ldots, F_S(x_n))). \tag{20}$$

where $Loss_{KD}$ and $Loss_{RKD}$ are distillation losses to penalize the difference between the teacher model and the student model. $F_T(x_i)$ and $F_S(x_i)$ represent two model functions, and λ^N is an n-tuple array collection of multiple sets of data, i.e., $\lambda^N = \{(x_i, x_j, \ldots) | (i \neq j \neq \ldots)\}$. $F_S(x_i)$ is the potential energy function, which is used to calculate the potential energy relationship between n-tuple arrays, and $Loss_{RKD}$ is the core part of the function. The teacher model transfers data to the student model through the potential energy to complete the knowledge transfer.

3. Experimental Results

3.1. Experimental Data Processing

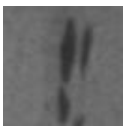
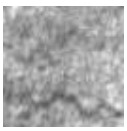
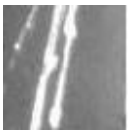
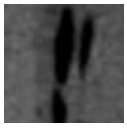
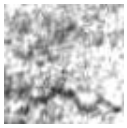
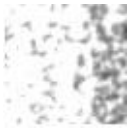
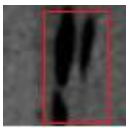
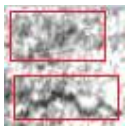
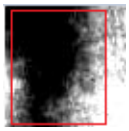
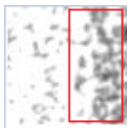
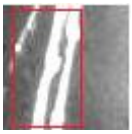
We have been influenced by some advanced work of strip defect detection and classification in industry 4.0 and manufacturing [34], references [35–37] and other deep learning technologies in industrial product defect detection and classification research work. Based on the actual environmental conditions of the industry, the relevant classification algorithm was designed, and in the process of designing the experiments, we fully considered how to meet the real-time performance of the shooting [38] and detection in the process of strip transmission, and how to deploy the operability of the algorithm application.

The experiment was carried out in this paper based on the NEU-DEU datasets. The image format of the dataset was adjusted to 64 × 64 RGB images, including 1600 images of various defects, namely: inclusion (In), crack (Cr), patches (Pa), pitted surface (Ps), rolled-in scale (Rs) and scratches (Sc). For all kinds of random samples, 1400 pieces were selected as training sets and 600 pieces were selected as testing sets. The label adopts the binary unique heat coding format. In this paper, the experiment was carried out under the Ubuntu system environment. The GPU model was an NVIDIA GTX 1080TI, and the PyTorch deep learning framework was used.

3.1.1. Image Preprocessing

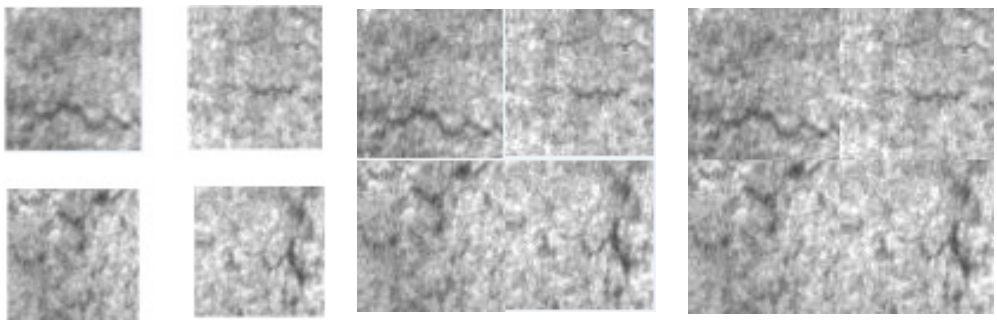
Low contrast and noise often appear in the transmission process of surface defect images of strip steel, so it is necessary to filter and enhance the original images. In this paper, the bilateral filter method [39] removes the low-frequency scanning noise in the image without destroying the edge information of the image, and a gray histogram is used to determine the gray transformation mapping curve [40]. The gray value of the image is adjusted and its contrast is stretched to enhance the image, and the OTSU is used to perform adaptive threshold segmentation. To facilitate observation, the defect areas with prominent features are marked with red wireframes, and the defect features are enhanced after preprocessing. The image preprocessing results are shown in Table 1.

Table 1. Image preprocessing.

	In	Cr	Pa	Ps	Rs	Sc
Original drawing						
Image enhancement						
Threshold segmentation						
Defect calibration						

3.1.2. Image Mosaicing and Fusion

Faced with the shortage of defect samples in a single production line, this paper randomly selects four 64 × 64 defect samples of the same type for splicing, and the image resolution after splicing is 128 × 128. To solve the problem that the gaps in the image edges interfere with the stitching effect, the median fuzzy method is used to process the gaps in the image stitching edges. Some original images of the samples are shown in Figure 8a. The result of image mosaicing is shown in Figure 8b, and the result after median blurring is shown in Figure 8c. Then, the resolution of the processed image is readjusted to 64 × 64 to form a new defect sample.



(**a**) Original images (**b**) Image mosaicing (**c**) Median blurring

Figure 8. Image mosaicing. (**a**) The image to be spliced. (**b**) Image mosaicing. (**c**) Median fuzzy image.

According to the analysis of Figure 8b, we can see that the stitching generated the same type of defect samples. After median fuzzy processing, the stitching gap in Figure 8c has been completely eliminated, and the image can be used as a new sample after resizing.

After image fusion processing in Figure 9, fine-grained texture information of two defect sample images can be collected, and it can be seen that the number of defects in the generated images has increased, but the types remain unchanged.

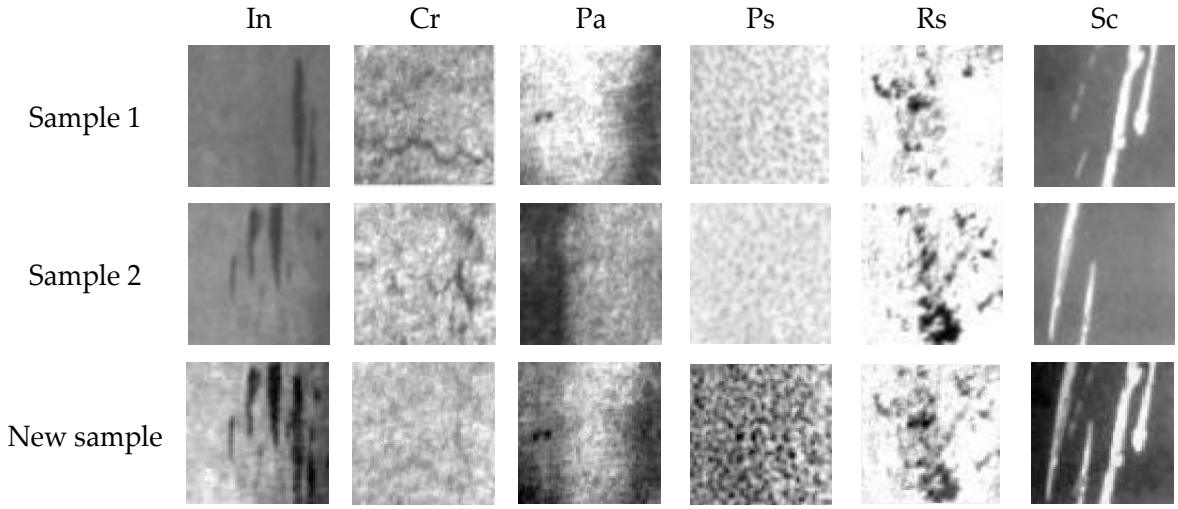


Figure 9. Image fusion.

3.1.3. Cycle GAN Data Enhancement Experiment

To evaluate the image enhancement effect of Cycle GAN data and verify the feasibility of cross-production line sample migration, a total of 1200 original images of various categories were acquired, and new defective samples were generated by image cross-domain conversion. The migration results of some defective samples are shown in Table 2

Table 2. Cycle GAN defect sample migration results.

	Comparison of Cycle GAN Defect Sample before and after Migration			
Sc original				
Sc generation				
In original				
In generation				
Rs original				
Rs generation				

In the example of Table 2, after the migration of Cycle GAN defect samples, samples that can be confused with real ones are generated. Sc, In, and Rs in the table are similar to the original images in terms of the size, position, shape, and texture details of the generated defects, which can be used to expand the sample data.

SARSN50 was used in the network, and the number of training and testing rounds was 50 epochs. The test accuracy results of each defect category before and after migration are shown in Figure 10 and shown on the right is the total test accuracy curve of defect samples before and after migration.

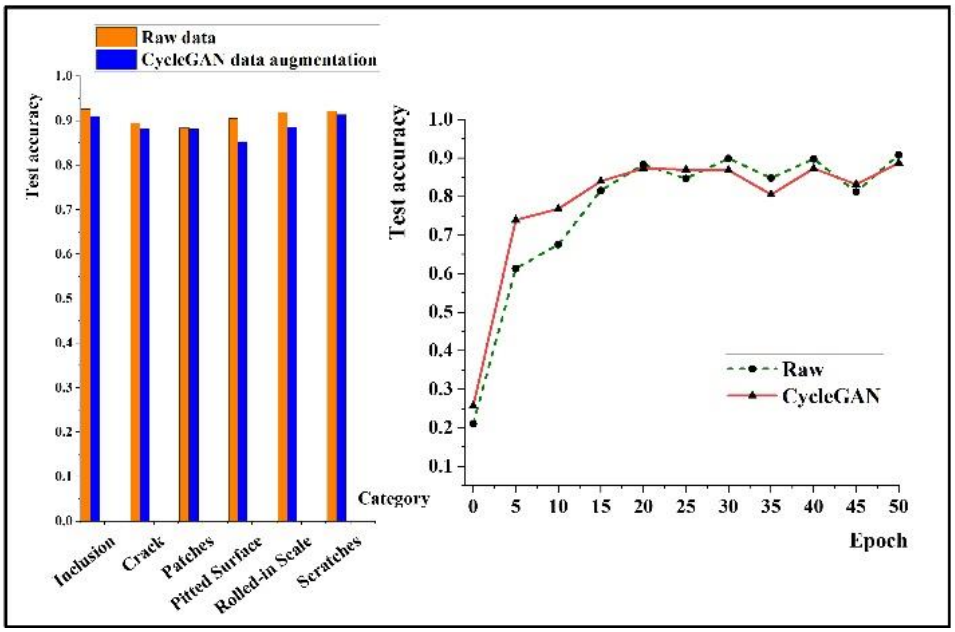


Figure 10. Cycle GAN data migration results.

The bar chart on the left side of Figure 10 illustrates the accuracy metrics on the test dataset after improving the defect sample migration strategy through Cycle GAN in comparison to the original data. It is evident that the blue bars, representing the test accuracy in six categories of steel strip defects, are very close to the yellow bars. This suggests that, with a slight sacrifice in accuracy, our approach can achieve higher classification prediction scores in completely unknown new domains. After training for 20 epochs, the test accuracy of the migrated image dataset begins to approach that of the original datasets. Under the condition that all parameters are the same, when 50 epochs are used, the test accuracy of the migrated image dataset reaches 88.7%, which is close to the test accuracy of the original dataset of 90.8%, which shows that the newly generated dataset has higher confidence after being enhanced by Cycle GAN and can initially meet the requirements of defect sample generation on new production lines in industrial fields. If the classification accuracy of migrated samples is further improved, a deeper SARSN101 network can be used.

3.2. Model Verification Experiment

3.2.1. Teacher Model

In this paper, SARSN34, 50, and 101 are used as teacher models in relational knowledge distillation. Except for the last fully connected layer of SARSN, all other layers are used as the feature extraction layer, and the fully connected layer is used as the classification layer. For the convenience of demonstration, the output characteristic map of the fourth layer of the teacher model training process is extracted, as shown in Figure 11a–f.

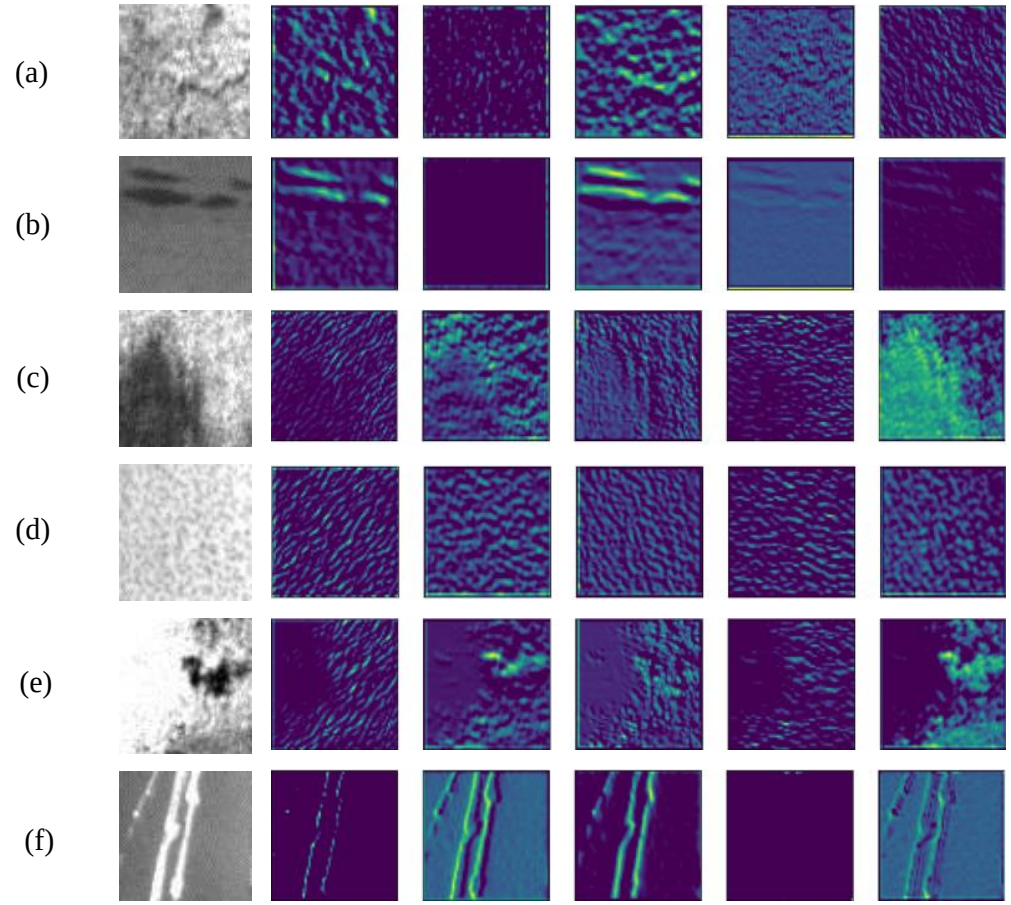
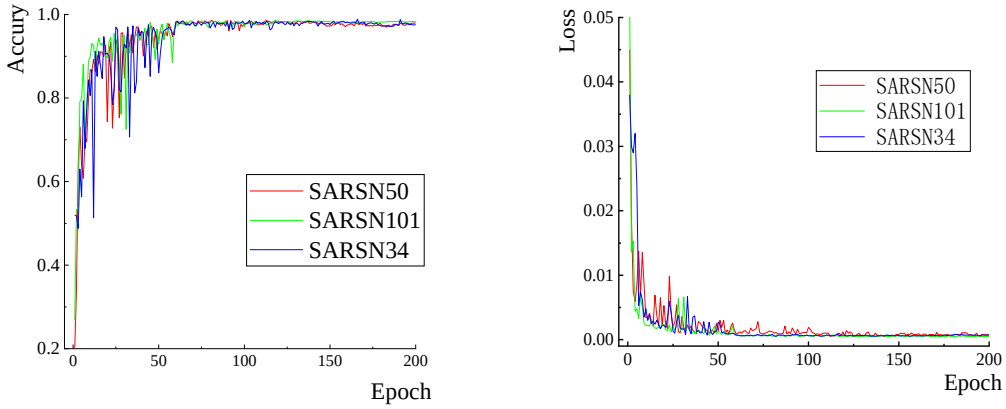


Figure 11. SARSN visualization of training process. (**a**) Crack. (**b**) Inclusion. (**c**) Patches. (**d**) Pitted surface. (**e**) Rolled-in scale. (**f**) Scratches.

Figure 11 shows that after the shallow feature information is extracted, the neural network can effectively divide the foreground area and the background area where the defect is located. For example, the green area in feature maps Figure 11a–c is identified as the defect area. Moreover, similar to small defects such as those of Figure 11a Cr and Figure 11d Ps, the feature extraction network can accurately capture all fine-grained information. It is shown that the teacher model can extract six categories of high-level features such as texture, shape, and edge.

The teacher model training process is set to 200 epochs, and the learning rate is 0.1 in the first 60 epochs, 0.02 in epochs 61–120, 0.004 in epochs 121~160, and 0.0008 in the final 40 epochs. During the initial stage of training, a warm upwarmup is used. The batch size is 8, the cross-entropy loss is used as the loss function, SGD algorithms are used to update the weight parameters, and the softmax function is used as the classification function. Figure 12a,b show the accuracy and loss curves of the teacher model training process, and the accuracy of teacher model training and testing is shown in Table 3. In the experimental process of training the teacher model, the expected number of training rounds is set to 200 epochs, and the training time is about 26 h. After about 120 epochs, the training iteration process of the model tends to be stable. The loss value of the training loss function also tends to be stable in the process of decline. At this time, the training iteration time is about 17 h.



(**a**) (**b**)

Figure 12. Teacher model training loss and accuracy curve. (**a**) Teacher model training accuracy curve. (**b**) Teacher model training loss curve.

Table 3. Teacher Model Training and Testing Accuracy.

Teacher Mode	Training Accuracy	Testing Accuracy
SARSN34	97.80%	96.85%
SARSN50	98.27%	97.14%
SARSN101	98.33%	97.83%

3.2.2. Student Model

To meet the real-time requirements of industrial field classification tasks, this paper chooses the lightweight network ResNet34 for the deployment of the student model of relational knowledge distillation. The training accuracy rates achieved when using SARSN34, 50, and 101 as teacher models to train the student model ResNet34 and only training ResNet34 are shown in Table 4, which shows that the student model has learned the relevant model generalization ability from the teacher model. With the deepening of the teacher model network, the training accuracy of the student model is also improved.

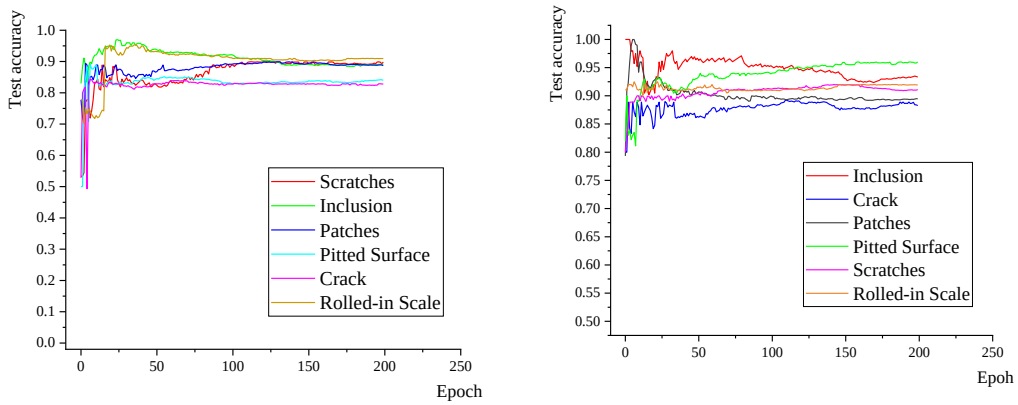
Table 4. Student model training accuracy.

Teacher Mode	Student Model	Training Accuracy
SARSN34	ResNet34	90.56%
SARSN50	ResNet34	92.31%
SARSN101	ResNet34	95.23%
_____	ResNet34	89.18%

According to the analysis of Table 4, the migration effect of the model with SARSN34 as the teacher is 90.56%, which is higher than 89.18% for the ResNet34 network in the same period, and the model efficiency is greatly improved. This shows that the student model has learned the relevant model generalization ability from the teacher model. With the deepening of the teacher model network, the training accuracy of the student model is also improved.

To verify the classification accuracy of the student model, aiming at six kinds of defects of strip steel, a total of 600 images were selected for testing, and the number of test rounds was 200 epochs. Taking the average accuracy of the current epochs as the test accuracy results, the average test accuracy values of SARSN34, SARSN50, and SARSN101 as the teacher model and ResNet34 as the student model were calculated to be 87.16%, 90.87%, and 94.40%, respectively. However, the average test accuracy rate of training the ResNet34 network alone is only 88.62%. The classification test accuracy results for six kinds of strip

steel defects are shown in Figure 13a,b. Figure 13a shows the test results of training the ResNet34 network independently without using relational knowledge distillation, and Figure 13b shows the test results of using SARSN50 as a teacher model and ResNet34 as a student model.



(**a**) (**b**)

Figure 13. Student model respective testing accuracy curve. (**a**) Original ResNet34 accuracy of each part, (**b**) Teacher-SARSN50 Student-ResNet34 accuracy of each part.

According to the analysis of Figure 13, the classification accuracy of six kinds of defects after knowledge distillation is generally improved, while the accuracy of six kinds of defects after RKD is improved by −1.46%, 2.25%, and 5.78% based on SARSN34, 50, and 101, respectively. This shows that the deeper the teacher model is, the more learnable parameters can be provided.

To further evaluate the testing effect of the student model obtained by distillation of relational knowledge, experiments were conducted on the 600 testing sets of strip steel defects. Each class contains 100 images, and the confusion matrix is shown in Figure 14, in which the abscissa represents the predicted value of each class of strip steel defect data, and the ordinate represents the true value. The classification accuracy P, the recall rate R, $F_1 \in [0, 1]$, and formulas of P, R and F1 are shown in Formula (21):

$$P = \frac{N_{TP}}{N_{TP} + N_{FP}}, R = \frac{N_{TP}}{N_{TP} + N_{FN}}, F_1 = \frac{2 \times P \times R}{P + R} \tag{21}$$

where NTP is the number of positive samples with correct prediction; NFP is the number of positive samples with prediction errors; NTN is the number of negative samples with correct prediction; and NFN is the number of negative samples with prediction errors. Combined with the data of the true values and predicted values of various defects in the confusion matrix, the classification accuracy, recall rate, and F1 value of various strip steel defects are calculated by Formula (21). The calculation results of the classification indicators are shown in Table 5.

Table 5. Evaluation of Classification Performance.

Defect Category	P	R	F_1
In	0.860	0.920	0.890
Pa	0.873	0.890	0.881
Ps	0.906	0.960	0.932
Sc	0.978	0.910	0.943
Rs	0.919	0.910	0.915
Cr	0.935	0.870	0.901

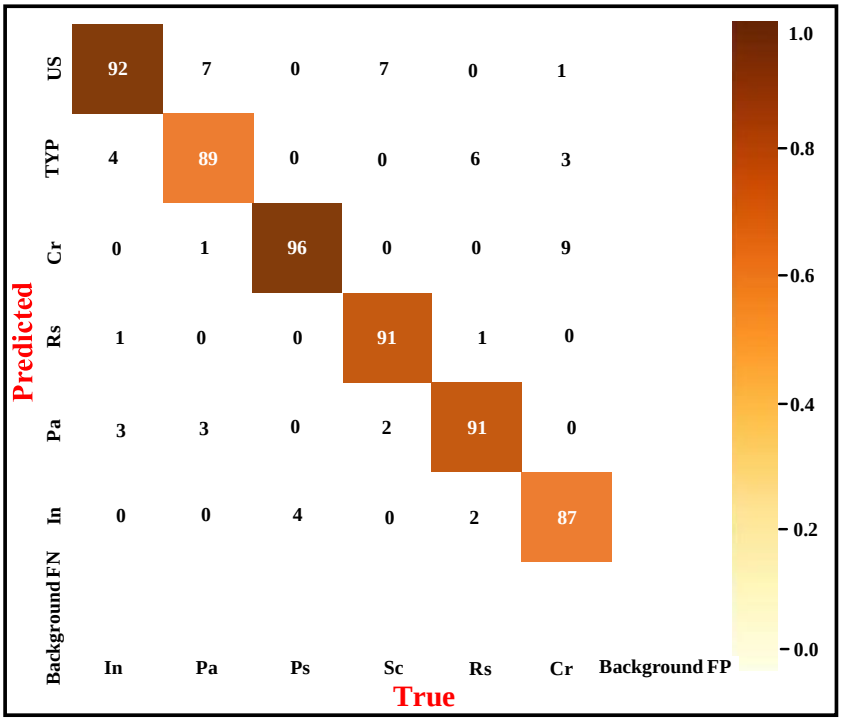
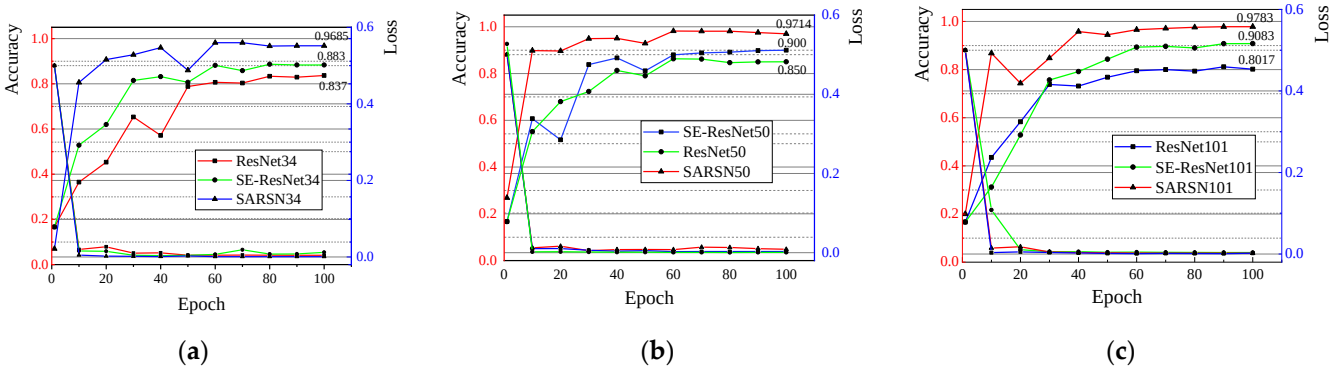


Figure 14. The confusion matrix.

Table 5 shows that the pitted surfaces and scratches exhibit excellent performance in terms of the F1 value, both of which have values above 93%. The F1 values of inclusions, patches and cracks have average performance, and the misjudgment rate is high, but the F1 values of the six types of defects are all approximately 90%, with good classification results. To further improve the classification accuracy of the six kinds of defect samples, the teacher model SARSN50 can be replaced by SARSN101 while maintaining the student model.

3.2.3. Model Comparative Experiment

SE-ResNet calibrates the feature response in the channel direction by stacking the squeezing and actuating blocks together. While combining the self-adaptive directional derivative threshold with the channel attention mechanism, SARSN can learn more useful fine-grained feature information in a specific threshold range. As shown in Figure 15, we selected ResNet34, 50, and 101, SE-ResNet34, 50, and 101 and SARSN34, 50, and 101, respectively. The training and testing conditions remain the same.



(**a**) (**b**) (**c**)

Figure 15. Comparison of classification performance of ResNet, SE-ResNet, and SARSN (**a**) 34 series, (**b**) 50 series, (**c**) 101 series.

From the comparison results in the figure, it can be found that under with the same network depth, the testing accuracy of the SE-ResNet network embedded with the SE block is better than that of the basic classification network ResNet, while the classification accuracy of SARSN proposed in this paper is the best of any model structure with any depth, among which the SARSN101 network reaches 97.83%. This shows that SARSN has the best effect in processing few-shot unbalanced data of strip steel.

To verify the classification performance of different network models for strip steel defects, this paper uses GoogLeNet, DenseNet, InceptionV3, SqueezeNet, MobileNet, ResNet101, ResNext101, SE-ResNet101, and Teacher-SARSN101 Student-ResNet34 networks for comparative experiments. The test accuracy curve of each model is shown in Figure 16. Moreover, the classification accuracy, parameters and single image classification algorithm time consumption of each network model are shown in Table 6.

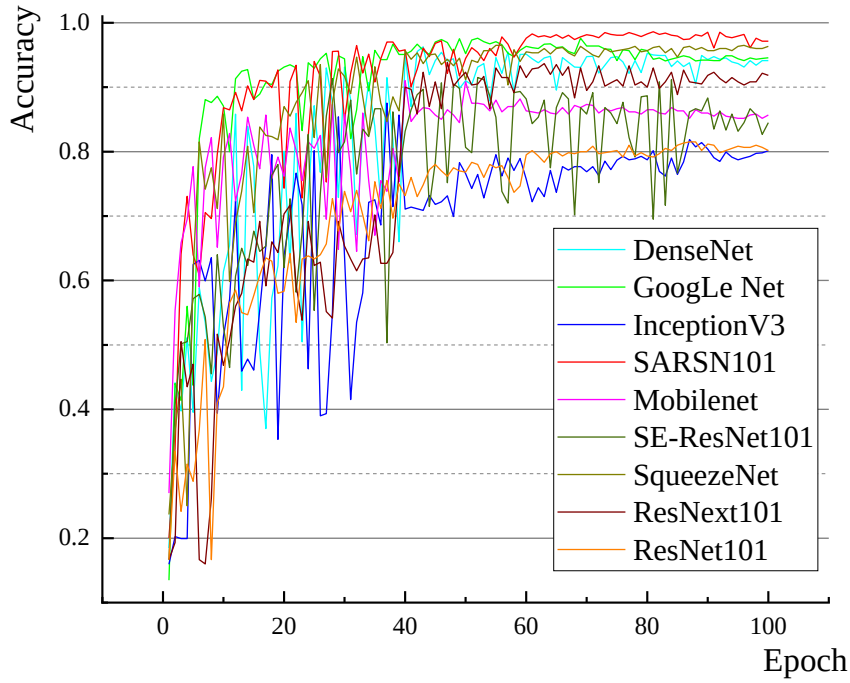


Figure 16. Accuracy curve of the classification test for each model.

Table 6. Comparison of the classification results of various models.

Model	Test Accuracy	Parameter Number	Time-Consumption/(ms)
GoogLeNet	0.9407	6,306,214	58
DenseNet	0.9413	6,952,198	96
InceptionV3	0.8011	22,125,542	248
SqueezeNet	0.9633	732,934	104
MobileNetV2	0.8567	3,219,078	7.8
ResNet101	0.8017	42,504,774	112
ResNext101	0.9167	14,788,772	67
SE-ResNet101	0.8451	47,527,764	187
SARSN101	0.9783	46,127,174	85
Teacher-SARSN101 Student-ResNet34	0.9440	11,171,910	51

It can be found that the test curve of the SARSN101 network rises gently, and its robustness is obviously better than that of other networks in the face of unbalanced datasets with few samples. Moreover, InceptionV3 is most obviously affected by the characteristics of the datasets, and its robustness is poor before 50 epochs. The testing accuracy of the SARSN101 network reaches 98.3%, which is much higher than that of the other models.

However, the test accuracy of the Teacher-SARSN 101 Student-ResNet 34 network using the relational knowledge distillation method is 94.4%, which is better than other network models in terms of the comprehensive performance of three indices.

To evaluate the method proposed in this paper more objectively, in addition to verifying that our proposed model has more advantages in accuracy and efficiency compared with other classification models, the following is also a comparison of different classification methods. After analysis, the texture information of six kinds of defects on the surface of strip steel is quite different, which is convenient for feature extraction. Four different feature extraction methods and classifier combinations were selected for comparison experiments with the methods proposed in this paper. In order to control the variance of the variables, all experiments selected identical datasets. The precision comparison results of each classification method are shown in Figure 17. The Gabor filter works on the principle of mutual modulation between the Gaussian kernel and sine wave, specializing in dealing with multi-scale and multi-directional texture features. We imported the features extracted by the Gabor filter into SVM and KNN classifiers, respectively to classify six kinds of defects. The traditional LBP (local binary pattern) algorithm has the advantages of rotation invariance and gray invariance and is better at describing local texture information. On the other hand, HOG (histogram of oriented gradient) has no rotation and scale invariance. The amount of calculations can be greatly reduced.

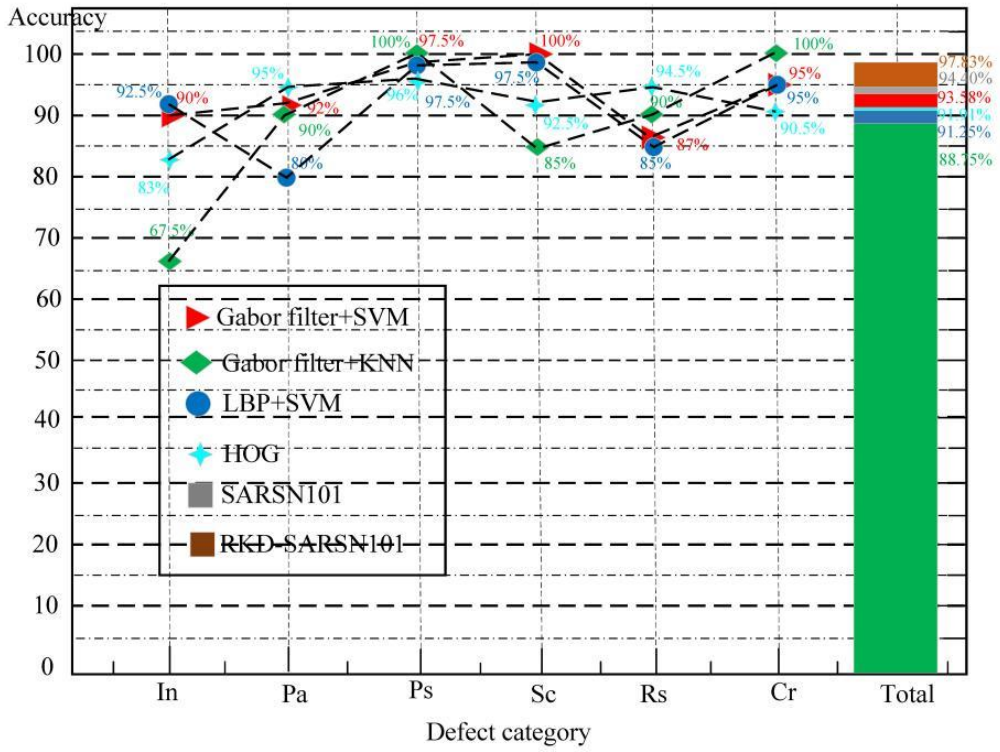


Figure 17. Comparison of the results of each classification algorithm.

From the classification results, it can be seen that the four kinds of algorithms have their own advantages in dealing with the classification effect of a single defect, among which, Gabor filter+SVM and Gabor filter+KNN methods achieve 100% classification accuracy for Sc, Ps and Cr, because the fine-grained texture features of Cr and Ps defects with different scales are obvious. However, in the overall classification accuracy of the six types of defects, the classification effect of SARSN101 and RKD-SARSN101 proposed in this paper is better.

3.2.4. Defect Calculation and Analysis

Deep neural networks have obvious advantages in feature extraction and fine-grained information acquisition, but large models take up significant computing space, and the defect features automatically extracted by neural networks are abstract, which is often not conducive to further analysis. The image moment has the characteristics of translation, rotation, and scale invariance. It is convenient to identify image features subjectively and simply. Among them, the 0-order moment represents the sum of the pixels, and the calculation formula is shown in Formula (22):

$$M_{00} = \sum_{I} \sum_{J} V(i,j). \tag{22}$$

The 1-order moment represents the product of the X-axis and Y-axis coordinates and their corresponding pixels in the rectangular coordinate system, and the calculation formula is shown in Formula (23):

$$M_{10} = \sum_{I} \sum_{J} i \cdot V(i,j); M_{01} = \sum_{I} \sum_{J} j \cdot V(i,j). \tag{23}$$

The centroid of the region to be studied is determined based on Formula (23) and shown in Formula (24):

$$x_c = \frac{M_{10}}{M_{00}}; y_c = \frac{M_{01}}{M_{00}}. \tag{24}$$

In this paper, using image processing technology and combining the characteristics of the 0-order moment (determination of the area of the target area) and the 1-order moment (determining the centroid of the target area), defect analysis is carried out on six kinds of samples with correct prediction output by the classification network, and the image processing results are shown in Figure 18.

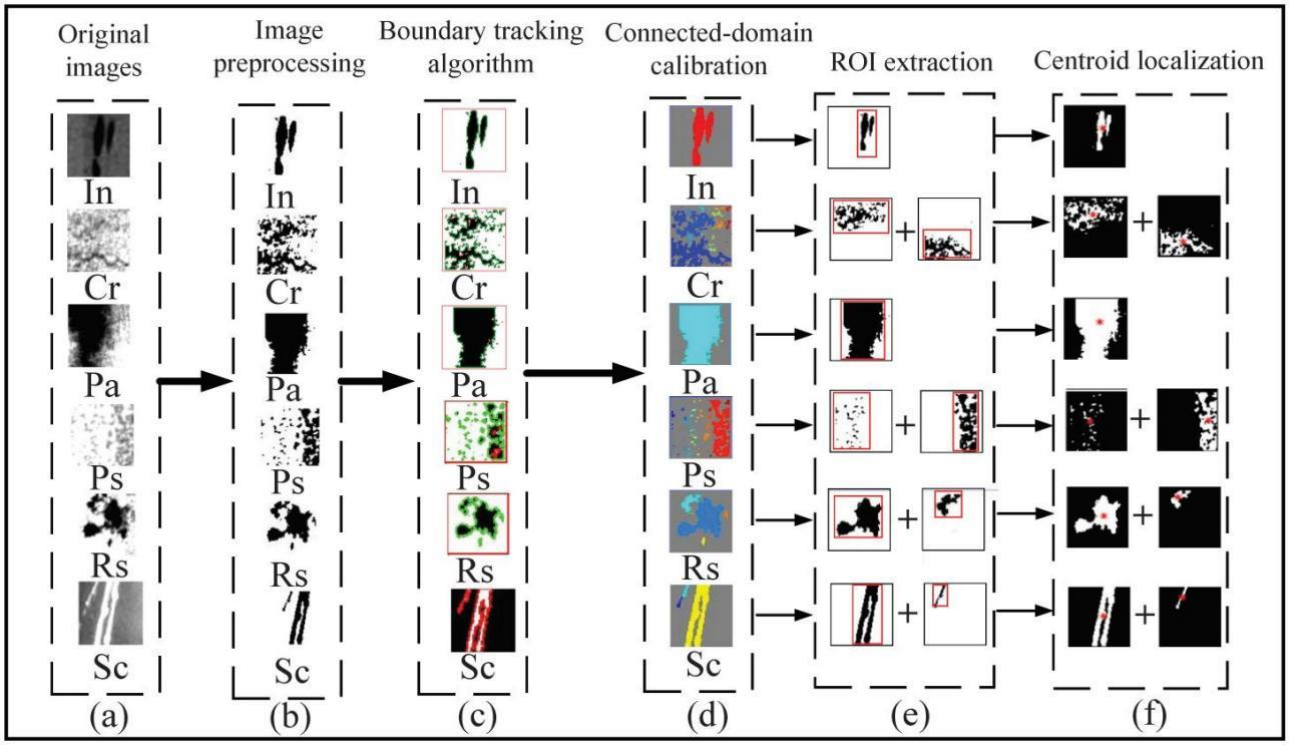


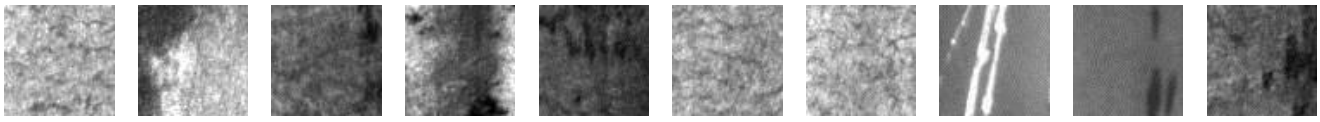
Figure 18. Image processing flow. (**a**) Original images. (**b**) Image preprocessing. (**c**) Boundary tracking algorithm. (**d**) Connected-domain calibration. (**e**) ROI extraction. (**f**) Centroid localization.

First, the original image (a) is preprocessed (b), the boundary tracing algorithm (c) is executed on the image after threshold segmentation and the perimeter of the defect area is calculated. Then, (b) the connected region is segmented. (d) It is convenient to extract the ROI of the defect feature. (e) The defect area is calculated. All the connected defect areas are marked with the same color, and the defects in each connected region are successfully segmented. Finally, the specific positions of all kinds of defects are determined by the centroid localization algorithm (f). The calculation results of the area and perimeter of various defects are shown in Table 7. We also specifically designed the proportion of defects and the area perimeter ratio (K) of defective parts to analyze the defects of each part.

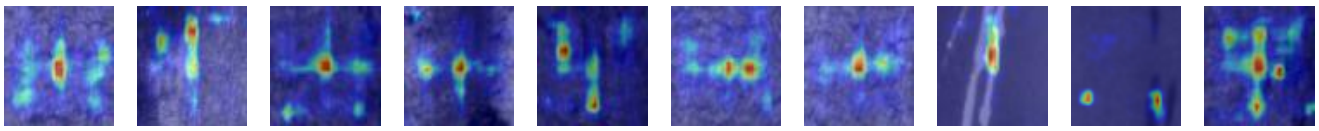
Table 7. Various defect evaluation indices.

Category	Area (Point)	Perimeter (Point)	Area Ratio (%)	Perimeter-Area Ratio (K)
Cr	1.646×10^3	475.911	40.18%	0.289
In	0.759×10^3	71.414	18.53%	0.094
Pa	2.332×10^3	113.314	56.93%	0.048
Ps	1.019×10^3	368.426	24.88%	0.362
Rs	1.159×10^3	213.012	28.30%	0.183
Sc	0.789×10^3	82.243	19.48%	0.103

In addition, in order to accurately locate the salient region of the defect being focused on by the neural network, we visualize the feature extraction results of the SARSN network by heat map, as shown in the results. Figure 19a is the original input image, and Figure 19b is the visualization result of the salient region, in which the dark color represents the area of focus of the neural network. From the results in the graph, it can be seen that the classification network can capture the distribution of some defects, indicating that our method can accurately identify the defect location through feature extraction.



(**a**) Original images



(**b**) Significant region extraction heat map visualization

Figure 19. Feature extraction of salient region heat map visualization results. (**a**) Original image. (**b**) Visualization results of the salient region.

According to the analysis in Table 7, the K values of Cr and Ps defects distributed in dense patches are all above 0.2, which indicates that these defects are characterized by small particles with a wide and dense distribution. The K values of In, Rs and Sc defects are distributed at approximately 0.1, which shows that this kind of defect has integrity and connectivity. However, the K value of Pa-type defects is less than 0.05, which indicates that Pa defects have the characteristics of a wide distribution area and high connectivity. Therefore, these characteristics can be used as the evaluation index of the defect grade of strip steel, which is of great significance in quantifying the defects of strip steel.

3.2.5. Design of the Grade Evaluation System for Surface Defects of Strip Steel

To effectively integrate the functions of each module of this algorithm and facilitate users in completing the task of defect classification without knowledge of the underlying

algorithm, we developed a set of strip steel surface defect rating systems for front-end deployment, with the interface information shown in Figure 20. The functions of this system include query data enhancement results, classification results, target detection results, and image processing results. Moreover, intermediate processes such as feature map visualization results and confusion matrices can be derived. The deployment of equipment, monitoring platforms, and local area networks needs to be determined according to the actual situation of the industrial site. In addition, the neural network training and testing processes of this system are all carried out offline, and the results are uploaded through the local area network.

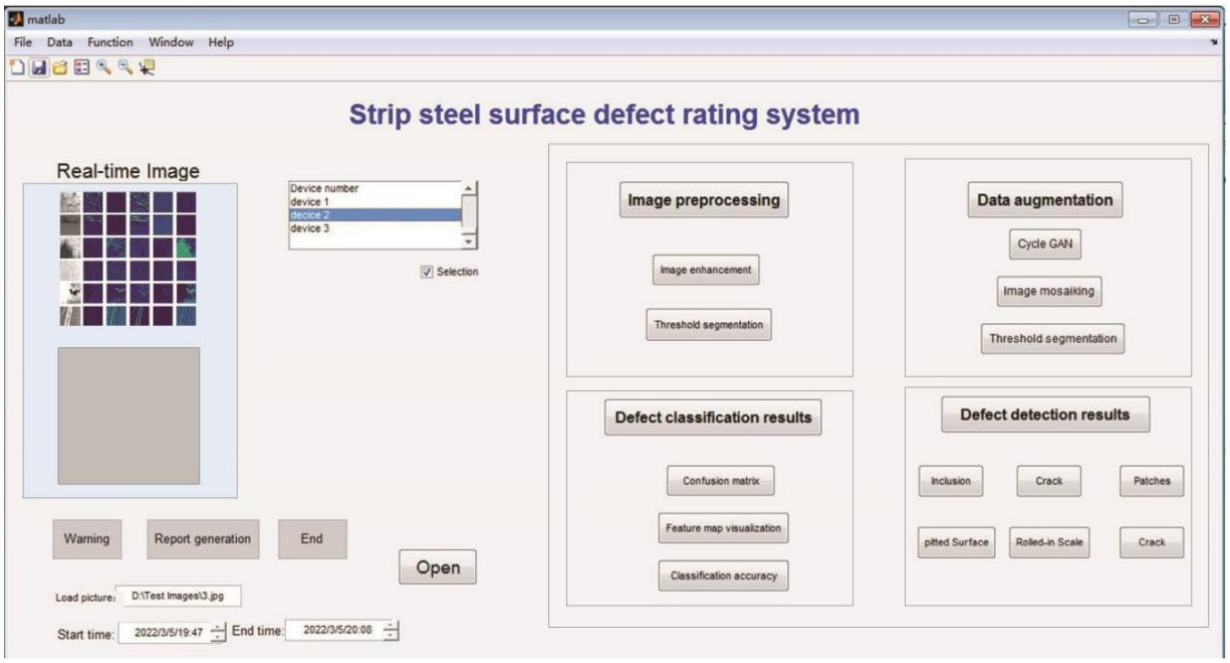


Figure 20. The evaluation system for surface defects of strip steel.

4. Conclusions

In this paper, a new relational knowledge distillation model framework network RKD-SARSN was proposed for the real-time classification of the sample data of the surface defects of strip steel. Through comparative experiments, the comprehensive performance of the proposed algorithms was found to be better than that of other classification algorithms, which can ensure classification accuracy and meet the real-time requirements of the classification of strip steel defects in the industrial field. The conclusions of the experiment are as follows:

- Using the Cycle GAN data enhancement method can realize cross-domain migration of defective samples of strip steel and solve the problem of few defective samples in a new production line;
- The introduction of the attention mechanism and self-adaptive directional derivative threshold in the SARSN model is the key to improving the classification accuracy of fine-grained defect images;
- In the training process of the model, the self-adaptive loss function balances the differences between classes through a separate class processing mechanism, which is helpful in solving the imbalance problem of strip steel defect samples;
- Structured relational knowledge distillation can transfer the generalization performance of large complex networks to small lightweight networks, reduce the complexity of model calculation and improve the efficiency of model deployment.

However, the method proposed in this paper may have limitations, which need to be ameliorated in future steps:

- The proposed self-adaptive loss function solves the problem of imbalanced samples between classes, but the sensitivity to highly imbalanced samples is average, and research in this area may need to be strengthened in the future;
- The method of structured relational knowledge distillation is outstanding in improving the detection efficiency of fine-grained image classification tasks, and it is beneficial to deploy it in industrial fields to solve practical problems. However, this method is rarely used in the fault diagnosis of mechanical vibration noise; thus, improving the fault diagnosis efficiency of vibration noise may be a future research direction.

Author Contributions: All authors contributed to the study conception and design. Z.S.: preliminary investigation, methodology, writing—original draft. X.H. and C.J.: verification, writing—review and editing, supervision. Y.Z. and L.Y.: project management and collecting documents, modifying formats, reference materials. All authors have read and agreed to the published version of the manuscript.

Funding: The state key laboratory open project of China National Heavy Machinery Research Institute. The Natural Science Basis Research Plan in Shaanxi Province of China (Program No.2022JQ-568). Scientific Research Program Funded by Shaanxi Provincial Education Department (Program No.21JK0661). Young Talent Fund of Association for Science and Technology in Shaanxi, China (Program No.20220133). Xi'an Science and Technology Plan Project (Program No.22GXFW0041). Shaanxi Province Innovative Talent Promotion Plant (2022KJXX-41). The Open Project of State Key Laboratory of Metal Extrusion and Forging Equipment Technology (No.S2208100.W03).

Data Availability Statement: Not applicable.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Bhattacharya, G.; Mandal, B.; Puhan, N.B. Interleaved Deep Artifacts-Aware Attention Mechanism for Concrete Structural Defect Classification. *IEEE Trans. Image Process.* **2021**, *30*, 6957–6969. [CrossRef] [PubMed]
2. Dong, X.; Taylor, C.J.; Cootes, T.F. Defect Detection and Classification by Training a Generic Convolutional Neural Network Encoder. *IEEE Trans. Signal Process.* **2020**, *68*, 6055–6069. [CrossRef]
3. Chagas, E.T.C.; Frery, A.C.; Rosso, O.A.; Ramos, H.S. Analysis and Classification of SAR Textures Using Information Theory. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **2021**, *14*, 663–675. [CrossRef]
4. Chu, M.X.; Feng, Y.; Yang, Y.H.; Deng, X. Multi-class classification method for steel surface defects with feature noise. *J. Iron Steel Res. Int.* **2021**, *28*, 303–315. [CrossRef]
5. Xu, H.; Yuan, H. An SVM-Based AdaBoost Cascade Classifier for Sonar Image. *IEEE Access* **2020**, *8*, 115857–115864. [CrossRef]
6. Ju, L.; Wang, X.; Zhao, X.; Lu, H.; Mahapatra, D.; Bonnington, P.; Ge, Z. Synergic Adversarial Label Learning for Grading Retinal Diseases via Knowledge Distillation and Multi-task Learning. *IEEE J. Biomed. Health Inform.* **2021**, *25*, 3709–3720. [CrossRef]
7. Guo, N.; Gu, K.; Qiao, J.; Bi, J. Improved deep CNNs based on Nonlinear Hybrid Attention Module for image classification. *Neural Netw.* **2021**, *140*, 158–166. [CrossRef]
8. Chiu, M.-C.; Chen, T.-M. Applying Data Augmentation and Mask R-CNN-Based Instance Segmentation Method for Mixed-Type Wafer Maps Defect Patterns Classification. *IEEE Trans. Semicond. Manuf.* **2021**, *34*, 455–463. [CrossRef]
9. Tu, Z.; Wu, S.; Kang, G.; Lin, J. Real-Time Defect Detection of Track Components: Considering Class Imbalance and Subtle Difference Between Classes. *IEEE Trans. Instrum. Meas.* **2021**, *70*, 1–12. [CrossRef]
10. Jiang, S.; Zhu, Y.; Liu, C.; Song, X.; Li, X.; Min, W. Data set Bias in Few-shot Image Recognition. *IEEE Trans. Pattern Anal. Mach. Intell.* **2021**, *45*, 229–246. [CrossRef]
11. Lv, N.; Ma, H.; Chen, C.; Pei, Q.; Zhou, Y.; Xiao, F.; Li, J. Remote Sensing Data Augmentation Through Adversarial Training. In Proceedings of the IGARSS 2020—2020 IEEE International Geoscience and Remote Sensing Symposium, Brussels, Belgium, 11–16 July 2021; pp. 2511–2514.
12. Lerner, B.; Yeshaya, J.; Koushnir, L. On the Classification of a Small Imbalanced Cytogenetic Image Database. *IEEE ACM Trans. Comput. Biol. Bioinform.* **2007**, *4*, 204–215. [CrossRef]
13. Jing, X.-Y.; Zhang, X.; Zhu, X.; Wu, F.; You, X.; Gao, Y.; Shan, S.; Yang, J.Y. Multiset Feature Learning for Highly Imbalanced Data Classification. *IEEE Trans. Pattern Anal. Mach. Intell.* **2021**, *43*, 139–156. [CrossRef] [PubMed]
14. Liang, X.; Zhang, Y.; Zhang, J. Attention Multisource Fusion-Based Deep Few-Shot Learning for Hyperspectral Image Classification. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **2021**, *14*, 8773–8788. [CrossRef]
15. Li, K.; Wang, X.; Ji, L. Application of Multi-Scale Feature Fusion and Deep Learning in Detection of Steel Strip Surface Defect. In Proceedings of the 2019 International Conference on Artificial Intelligence and Advanced Manufacturing (AIAM), Dublin, Ireland, 16–18 October 2019; pp. 656–661.

16. Wang, Z.; Wang, J.; Chen, S. Fault Location of Strip Steel Surface Quality Defects on Hot-Rolling Production Line Based on Information Fusion of Historical Cases and Process Data. *IEEE Access* **2020**, *8*, 171240–171251. [CrossRef]
17. Goodfellow, I.J.; Pouget-Abadie, J.; Mirza, M.; Xu, B.; Warde-Farley, D.; Ozair, S.; Courville, A.; Bengio, Y. Generative Adversarial Networks. *Adv. Neural Inf. Process. Syst.* **2014**, *3*, 2672–2680. [CrossRef]
18. Xia, K.; Yin, H.; Qian, P.; Jiang, Y.; Wang, S. Liver Semantic Segmentation Algorithm Based on Improved Deep Adversarial Networks in combination of Weighted Loss Function on Abdominal CT Images. *IEEE Access* **2019**, *7*, 96349–96358. [CrossRef]
19. Creswell, A.; White, T.; Dumoulin, V.; Arulkumaran, K.; Sengupta, B.; Bharath, A.A. Generative Adversarial Networks: An Overview. *IEEE Signal Process. Mag.* **2018**, *35*, 53–65. [CrossRef]
20. He, K.; Zhang, X.; Ren, S.; Sun, J. Deep Residual Learning for Image Recognition. In Proceedings of the 2016 IEEE conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, USA, 27–30 June 2016; pp. 770–778.
21. Xie, S.; Girshick, R.; Dollár, P.; Tu, Z.; He, K. Aggregated Residual Transformations for Deep Neural Networks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, USA, 27–30 June 2016; pp. 1492–1500.
22. Zhao, M.; Zhong, S.; Fu, X.; Tang, B.; Pecht, M. Deep Residual Shrinkage Networks for Fault Diagnosis. *IEEE Trans. Ind. Inform.* **2020**, *16*, 4681–4690. [CrossRef]
23. Jie, H.; Li, S.; Gang, S.; Albanie, S. Squeeze-and-Excitation Networks. *IEEE Trans. Pattern Anal. Mach. Intell.* **2017**, *42*, 2011–2023.
24. Verma, A.; Sharma, M.; Hebbalaguppe, R.; Hassan, E.; Vig, L. Automatic Container Code Recognition via Spatial Transformer Networks and Connected Component Region Proposals. In Proceedings of the IEEE International Conference on Machine Learning and Applications, Anaheim, CA, USA, 18–20 December 2016; pp. 728–733.
25. Woo, S.; Park, J.; Lee, J.Y.; Kweon, I.S. CBAM: Convolutional Block Attention Module. In Proceedings of the European Conference on Computer Vision, Munich, Germany, 8–14 September 2018; pp. 3–19.
26. Sun, C.; Ma, M.; Zhao, Z.; Chen, X. Sparse deep stacking network for fault diagnosis of motor. *IEEE Trans. Ind. Inf.* **2018**, *14*, 3261–3270. [CrossRef]
27. Chen, Y.; Bai, Y.; Zhang, W.; Mei, T. Destruction and Construction Learning for Fine-Grained Image Recognition. In Proceedings of the 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Long Beach, CA, USA, 15–20 June 2019; pp. 5157–5166.
28. Leng, R.; Zhou, W. Optimization Research and Application of Unbalanced Data Set Multi-classification Algorithm. In Proceedings of the International Conference on Intelligent Human-Machine Systems and Cybernetics, Hangzhou, China, 27–28 August 2016; pp. 39–42.
29. Raj, V.; Magg, S.; Wermter, S. Towards effective classification of imbalanced data with convolutional neural networks. In Proceedings of the IAPR Workshop on Artificial Neural Networks in Pattern Recognition, Ulm, Germany, 28–30 September 2016; pp. 150–162.
30. Elhanashi, A.; Gasmi, K.; Begni, A.; Dini, P.; Zheng, Q.; Saponara, S. Machine Learning Techniques for Anomaly-Based Detection System on CSE-CIC-IDS2018 Dataset. In *Applications in Electronics Pervading Industry, Environment and Society. ApplePies 2022*; Berta, R., De Gloria, A., Eds.; Lecture Notes in Electrical Engineering; Springer: Cham, Switzerland, 2023; Volume 1036. [CrossRef]
31. Anwary, A.R.; Yu, H.; Vassallo, M. Gait Evaluation Using Procrustes and Euclidean Distance Matrix Analysis. *IEEE J. Biomed. Health Inform.* **2019**, *23*, 2021–2029. [CrossRef] [PubMed]
32. Hinton, G.; Vinyals, O.; Dean, J. Distilling the Knowledge in a Neural Network. *Comput. Sci.* **2015**, *14*, 38–39.
33. Park, W.; Kim, D.; Lu, Y.; Cho, M. Relational Knowledge Distillation. In Proceedings of the 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Long Beach, CA, USA, 15–20 June 2019; pp. 3967–3976.
34. Angelopoulos, A.; Michailidis, E.T.; Nomikos, N.; Trakadas, P.; Hatziefremidis, A.; Voliotis, S.; Zahariadis, T. Tackling faults in the industry 4.0 era—A survey of machine-learning solutions and key aspects. *Sensors* **2019**, *20*, 109. [CrossRef] [PubMed]
35. Hao, R.; Lu, B.; Cheng, Y.; Li, X.; Huang, B. A steel surface defect inspection approach towards smart industrial monitoring. *J. Intell. Manuf.* **2021**, *32*, 1833–1843. [CrossRef]
36. Tang, B.; Chen, L.; Sun, W.; Lin, Z.K. Review of surface defect detection of steel products based on machine vision. *IET Image Process.* **2023**, *17*, 303–322. [CrossRef]
37. Czimmermann, T.; Ciuti, G.; Milazzo, M.; Chiurazzi, M.; Roccella, S.; Oddo, C.M.; Dario, P. Visual-based defect detection and classification approaches for industrial applications—A survey. *Sensors* **2020**, *20*, 1459. [CrossRef]
38. Sampath, V.; Maurtua, I.; Martín, J.J.A.; Rivera, A.; Molina, J.; Gutierrez, A. Attention guided multi-task learning for surface defect identification. *IEEE Trans. Ind. Inform.* **2023**, *19*, 9713–9721. [CrossRef]
39. Nnolim, U.A. Multi-Scale Fractional Tonal Correction Bilateral Filter-Based Hazy Image Enhancement. *Int. J. Image Graph.* **2020**, *20*, 2050010. [CrossRef]
40. Tian, F.; Gao, Y.; Fang, Z.; Gu, J. Automatic coronary artery segmentation algorithm based on deep learning and digital image processing. *Appl. Intell.* **2021**, *51*, 8881–8895. [CrossRef]