Article

Assembly Quality Detection Based on Class-Imbalanced Semi-Supervised Learning

Zichen Lu [1], Jiabin Jiang [2], Pin Cao [2] and Yongying Yang [1,*]

1   Key Laboratory of Modern Optical Instrumentation, Department of Optical Engineering, Zhejiang University, 38 Zheda Road, Hangzhou 310027, China; sakumashirayuki@gmail.com
2   Zernike Optics Co., Ltd., 10 Shengang Road, Hangzhou 310027, China; jiabinjiang@zju.edu.cn (J.J.); coolercp@126.com (P.C.)
*   Correspondence: chuyyy@zju.edu.cn; Tel.: +86-0571-86141072

Abstract: Due to the imperfect assembly process, the unqualified assembly of a missing gasket or lead seal will affect the product's performance and possibly cause safety accidents. Machine vision method based on deep learning has been widely used in quality inspection. Semi-supervised learning (SSL) has been applied in training deep learning models to reduce the burden of data annotation. The dataset obtained from the production line tends to be class-imbalanced because the assemblies are qualified in most cases. However, most SSL methods suffer from lower performance in class-imbalanced datasets. Therefore, we propose a new semi-supervised algorithm that achieves high classification accuracy on the class-imbalanced assembly dataset with limited labeled data. Based on the mean teacher algorithm, the proposed algorithm uses certainty to select reliable teacher predictions for student learning dynamically, and loss functions are modified to improve the model's robustness against class imbalance. Results show that when only 10% of the total data are labeled, and the imbalance rate is 5.3, the proposed method can improve the accuracy from 85.34% to 93.67% compared to supervised learning. When the amount of annotated data accounts for 20%, the accuracy can reach 98.83%.

Keywords: intelligent quality inspection; semi-supervised learning; imbalanced data; mean teacher

1. Introduction

Screw fasteners are simple in construction and easy to operate. They are widely used in the mechanical structure as the crucial part of large equipment, such as aero-engine, high-speed railways, production machinery, wind turbines, air conditioning systems, and elevator cranes. The assembly quality should be carefully inspected because it determines the mechanical properties and safety of products. Screw–gasket–seal is a typical connector assembly structure. The gasket is used to protect the surface of the connector from screw abrasion. The seal can pass through several structures to form a closed loop, preventing the structure from being loose. The images of qualified assembly samples are shown in Figure 1(a1–a4), with a lead threading through two parallel holes of the hexagon screw and a gasket placed between the two mating surfaces to increase friction. Unqualified assemblies (Figure 1(b1–b4)) without gaskets or lead seals have hidden safety hazards, which may cause inestimable loss of life and property.

Therefore, it is crucial to detect whether such assemblies are qualified effectively. With the advent of industry 4.0 and the continuous development of machine learning technology, the manufacturing industry has an increasing demand for automatic and intelligent production [1,2]. Both industrial practitioners and academic researchers are exploring intelligent detection methods to replace manual inspection. Common noncontact inspection methods include 3D scanning [3–5] and machine vision. The 3D scanning methods can provide the precision position of the component to assist the industrial robots. Machine vision detection methods use the camera to capture images from production lines and then design

algorithms to complete the extraction and analysis of image information. It can provide objective, high-speed measurement and good reliability with a simple system and strong adaptability. Before the rise of deep learning algorithms, there have been many studies based on low-level visual features. For example, in [6], Hoff transform was applied to the bolt-loosening detection in the connection of wind turbine tower segments; Liu et al. [2] used the gradient coded co-occurrence matrix (GCCM) to inspect the missing of bogie block key on freight cars. Due to its data-driven nature, the fast-developing deep learning algorithms can extract knowledge from historical data, reducing the dependence on expert domain knowledge and avoiding artificial design of visual features. They have been widely used in the detection of key components of high-speed trains [7–10], fault diagnosis [11–13], high-voltage transmission line detection [14–16], and other industrial applications.
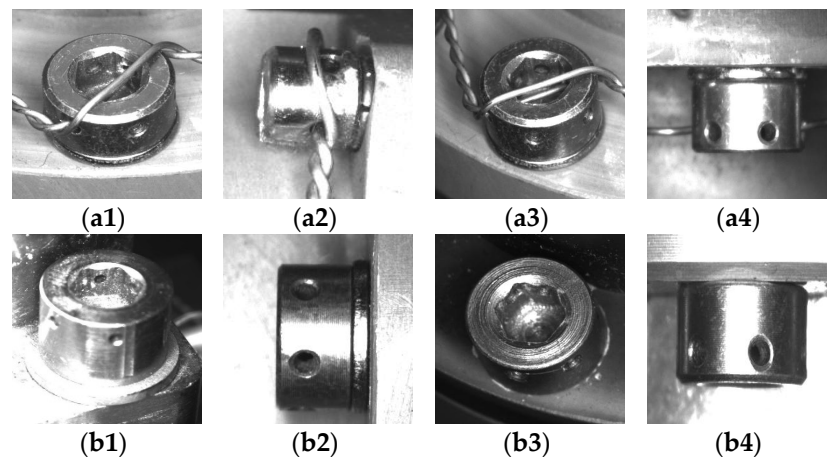

Figure 1. Assembly sample images: (**a1**–**a4**) Qualified assemblies; (**b1**,**b2**) Samples without lead sealing; (**b3**,**b4**) Samples without lead sealing and gasket.

The high generalization performance of deep learning relies on a large amount of labeled data. However, massive manpower and repeated labor will be needed to collect and annotate tens of thousands of data, and the sample number of the unqualified class is often low, resulting in the class imbalance of the training set. These factors will significantly affect the neural network model. Few-shot learning, generative methods, and semi-supervised learning have been presented to alleviate these problems. Few-shot learning is one method that fully utilizes a small number of labeled data [17]. It is suitable for tasks where a variety of prior knowledge is available, including supervised data from other domains and modalities. Wang et al. [11] used a classification algorithm based on the similarity of sample pairs in intelligent bearing fault diagnosis, which included feature learning and metric learning modules. The feature learning module used twin neural networks to extract features from the sample pair separately, and the metric learning module was used to predict the similarity of the sample pair. The classification was conducted according to the similarity between the test sample and the labeled sample. To avoid artificially designing similarity measure function, [18] introduced the meta-learning method into the metric learning module in machine fault diagnosis to learn distance function adaptively.

Some studies attempt to generate data artificially. A common generating method is to use the generative adversarial network (GAN). Two models are trained simultaneously: one generative model G capturing the data distribution, and a discriminant model D estimating the probability that samples come from real data rather than G [19]. One study [20] used a generator to generate faulty mode data in compressor fault detection by minimizing the cluster center distance between the real and the generated data. In [21], the generative model was also applied to intelligent fault diagnosis of rotating machinery, where feature differences obtained by a feature extractor were minimized to obtain high-quality generated data. Compared with the classification network, the generative

model requires an additional design of the generator structure. Moreover, it needs careful adjustment of the hyper-parameters and more time in optimizing the network structure to prevent divergence.

When unlabeled samples are available in large quantities, semi-supervised learning (SSL) is a promising approach. There are two main kinds of semi-supervised learning methods: unsupervised preprocessing and perturbation-based method. Unsupervised preprocessing methods usually extract features from unlabeled data. For example, in detecting freight car plate bolts, [22] used unlabeled data to pre-train the stacked auto-encoder and then assigned its weights to the classification network. Zhang et al. [12] combined the training stages of auto-encoder and classifier and used two identical encoder networks to process labeled and unlabeled data for bearing fault diagnosis simultaneously. This integrated training method can achieve higher classifier accuracy.

The consistency regularization method based on perturbation is used to find the smooth manifold where the dataset lies by leveraging the unlabeled data [23]. It is assumed that similar samples have similar labels in the dense data space. Consistency regularization methods usually add perturbation to input data, network structure, and training mode and constrain the probability distribution of the model's output to remain unchanged. They do not depend on any intermediate steps or pre-trained supervised learners and generally extend the existing supervised loss function to contain the unlabeled data. In the Π model [24], each sample of the labeled and unlabeled dataset is propagated twice in every epoch, and perturbations are introduced by random noise of the input data and network dropout. To reduce the computing burden, the temporal ensembling method [25] replaces one forward propagation with the exponential moving average (EMA) of earlier predictions. The mean teacher (MT) method [26] directly applies the EMA to the regular network parameters to obtain another Teacher model. Compared with Temporal Ensembling, it can get more accurate predictions. There are few studies on semi-supervised learning methods for detecting unqualified assemblies or other products. The academic studies of consistency regularization methods almost assume that the distribution of instances in each class is balanced [19]. However, they have difficulty ensuring ideal performance on imbalanced data, especially in minority classes. Sometimes they are even worse than supervised learning methods [27].

To address the issues above, this article introduces a semi-supervised learning algorithm to detect unqualified assembly samples. It can achieve an accuracy of 93.67% when the labeled fraction of the training dataset is 10% and the imbalance rate is 5.3. This algorithm improves the mean teacher algorithm and makes up for the deficiency of the semi-supervised learning method in the class-imbalanced scenarios. Firstly, the certainty values of teacher predictions are measured, and the teacher predictions with high certainty are selected for the consistency constraint. Then, label-distribution-aware margin loss (LDAM Loss) [28] is applied to the labeled data training of the student model to enhance the robustness against class imbalance under supervised learning, and compression consistency loss (CCL) is adopted to prevent decision boundaries from skewing into the minority class regions.

This paper is organized as follows: Section 2 introduces the proposed semi-supervised learning algorithm in detail. In Section 3, the assembly dataset and comparative experiments between the proposed method and other existing methods are described, and the results are discussed. Finally, Section 4 presents the main conclusions.

2. Class-Imbalanced Semi-Supervised Learning

Assume that the training set D contains N samples, N_l of which are labeled. Let $D_l = \{(x_i, y_i) : y_i \in (1, \ldots, C)\}_{i=1}^{N_l}$ represents the labeled training set, where the training sample is denoted as x_i, y_i is the corresponding one-hot label, and C is the number of classes. D_u is the unlabeled training set. The mini-batch size is B at each iteration, consisting of b_l labeled samples and $b_u = B - b_l$ unlabeled samples. In this section, the proposed model structure and the training strategy are introduced in detail.

2.1. Model Framework for Assembly Quality Detection

The method adopted in this paper is based on the mean teacher algorithm, which includes a teacher model and a student model with the same network structure. The overall algorithm is shown in Figure 2. The models use the DenseNet121 [29] structure, famous for achieving high performance while reducing the scale of parameters. As shown in Figure 2, the DenseNet121 includes four dense blocks, each followed by a transition layer. The transition layer consists of one 1×1 convolutional layer and one 2×2 pooling layer. The dense block is composed of multiple dense layers, and the output of each dense layer is connected to other layers in a feedforward manner to prevent the disappearance of features. The dense block and dense layer are illustrated in Figure 3, and each dense layer is a sequence of BN-ReLU-1×1 Conv-BN-ReLU-3×3 Conv.
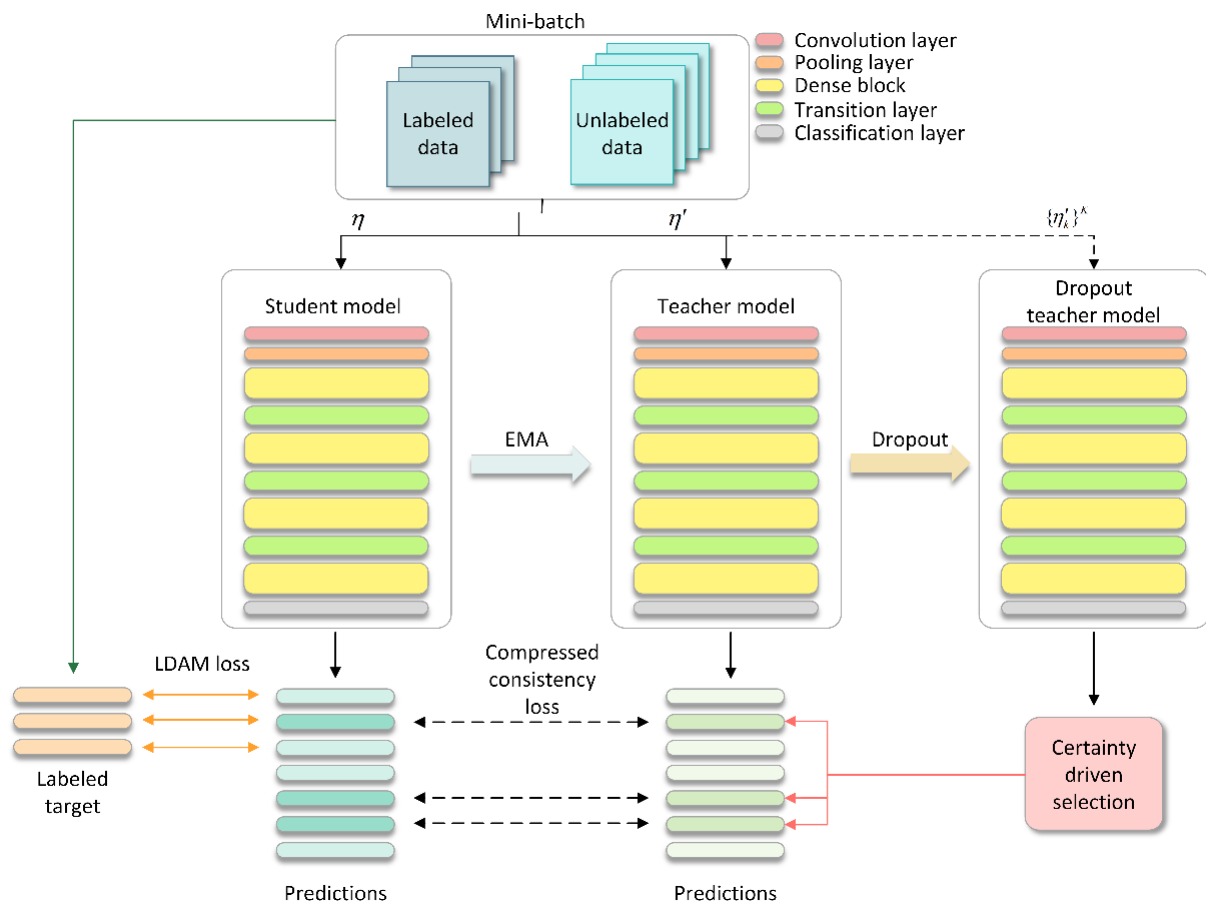

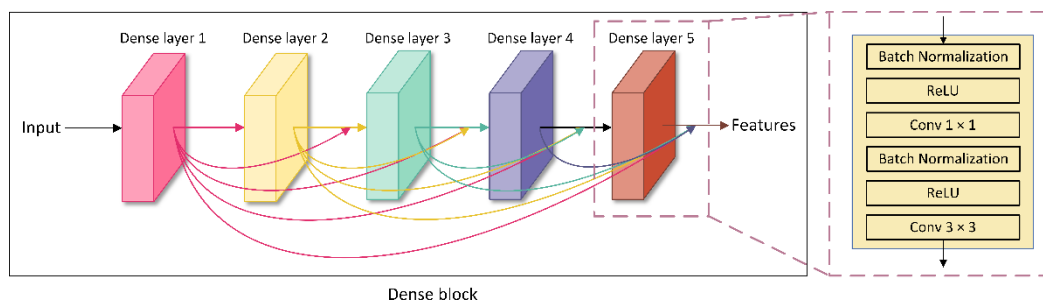**Figure 2.** Class-imbalanced semi-supervised learning framework.

Figure 3. Dense block composed of five Dense layers.

Inspired by knowledge distillation, the mean teacher (MT) method [26] uses the teacher-student structure. The weights of the teacher are the exponential moving average of the weights of the student. The mean teacher algorithm introduces perturbations in the model weights and input data and encouraging the predictions to remain the same. Defining θ^s as the weights of the student model, then the corresponding weight θ^t of the teacher model is

$$\theta^t \leftarrow \mu\theta^t + (1 - \mu)\theta^s, \tag{1}$$

$$\mu = \min(1 - \frac{1}{iter}, \mu_0), \tag{2}$$

where μ is the smoothing coefficient hyper-parameter, *iter* is the global iteration step and μ_0 is the maximum value of μ. At the early training period, μ is small, therefore the teacher is rapidly updated by the new student weights. In the later training period, when μ reaches μ_0, the teacher will have a longer memory since the improvement of the student is slow down. As the perturbations to the input, random noise enhancement (η, η') is applied to the original sample x_i before input to the models, so the predictions of the student and the teacher are $\hat{y}_i^s = f(x_i, \theta^s, \eta)$ and $\hat{y}_i^t = f(x_i, \theta^t, \eta')$. The usual mean teacher approach uses the mean square error (MSE) as the consistency regularization loss (CRL) to minimize the Euclidean distance between the teacher prediction and the student prediction.

$$L_{CRL}(\hat{y}_i^s, \hat{y}_i^t) = \left\| \hat{y}_i^s - \hat{y}_i^t \right\|^2, \tag{3}$$

Not all teacher predictions are reliable, and consistent constraints on the unreliable predictions will damage model performance. To make the student model dynamically select reliable predictions from the teacher, this paper adopts the certainty driven mechanism, which is explained in detail in Section 2.2. At each iteration, m samples with reliable teacher predictions are chosen from the mini-batch to form the subset M, and the consistent constraint is computed on M. Then, the cost function is the sum of the consistency regularization loss of each sample.

$$J_{consistency} = \sum_{x_i \in M} L_{CRL}(\hat{y}_i^s, \hat{y}_i^t), \tag{4}$$

To improve the robustness of MT to class imbalance, LDAM loss is applied to b_l labeled samples in the mini-batch to enhance the accuracy of the student model in the supervised training. Moreover, the compression consistency loss (CCL) is introduced to weaken the decision boundary smoothing effect of samples predicted to be the majority class. Details of the class-imbalanced loss functions are described in Section 2.3. Finally, the semi-supervised cost function of the mini-batch is

$$\begin{aligned} J &= J_{classification} + J_{consistency} \\ &= \sum_{i=1}^{b_l} L_{LDAM}(\hat{y}_i^s, y_i) + \sum_{x_i \in M} L_{CCL}(\hat{y}_i^s, \hat{y}_i^t) \end{aligned} \tag{5}$$

The pseudo code in Algorithm 1 presents the whole process of the proposed method.

Algorithm 1 Training of the proposed method

Input: $D_l, D_u, B, b_l, \beta, K$
Initialization: θ^s, θ^t
for iter = 1 T **do**

$\quad$ Sample batches $\{x_i\}_{i=1}^B, \{y_i\}_{i=1}^{b_l}$
$\quad$ Student prediction $\hat{y}_i^s \leftarrow f(x_i, \theta^s, \eta)$
$\quad$ Dropout samplings $\tilde{y}_{i,k}^t \leftarrow f(x_i, \theta_k^t, \eta'_k), k = 1, \ldots, K$
$\quad$ Teacher prediction $\hat{y}_i^t \leftarrow f(x_i, \theta^t, \eta')$
$\quad$ Compute uncertainty $U(x_i) \leftarrow Var(\tilde{y}_{i,1}^t, \ldots, \tilde{y}_{i,K}^t)$
$\quad$ Rearrange in ascending order $\{P_1, \ldots, P_B\} \leftarrow \{U(x_1), \ldots, U(x_B)\}$
$\quad$ $m \leftarrow \min(\beta e, B)$
$\quad$ $M \leftarrow \{P_1, \ldots, P_m\}$
$\quad$ $J \leftarrow \sum_{i=1}^{b_l} L_{\text{LDAM}}(\hat{y}_i^s, y_i) + \sum_{x_i \in M} L_{\text{CCL}}(\hat{y}_i^s, \hat{y}_i^t)$
$\quad$ Update the student model $\theta^s \leftarrow \theta'^s \leftarrow \theta^s - \alpha \nabla \theta^s$
$\quad$ Update the teacher model $\theta^t \leftarrow \theta'^t \leftarrow \mu\theta^t + (1-\mu)\theta^s$

end

2.2. Certainty Driven Selection

The deficiency of existing perturbation-based semi-supervised learning methods is that all the outputs are regularized without exception. A large part of the outputs can be unreliable due to the confirmation bias [26]. Confirmation bias results from incorrect predictions of unlabeled data used in subsequent training, increasing the confidence of wrong predictions and making the model resist new changes. In this case, maintaining the consistency regularization will result in the student model converging incorrectly. In the absence of labeled targets as supervision, evaluating the certainty of teacher predictions, and filtering out low-certainty samples is necessary to ensure that the consistency constraints only apply to high-certainty samples.

The certainty driven selection method is shown in Figure 4. It is assumed that the teacher network has H layers, with the parameters set $\theta^t = \{\Phi_h\}_{h=1}^H$ determined by limited random variables, and Φ_h represents the parameters of layer h. For sample x_i, the predicted distribution $q(\hat{y}_i^t | x_i)$ of the teacher is approximated as [30]

$$q(\hat{y}_i^t | x_i) = \int p(\hat{y}_i^t | x_i, \theta^t) q(\theta^t) d\theta^t, \tag{6}$$

where $p(\hat{y}_i^t | x_i, \theta^t)$ is the prediction probability based on the input data x_i and model parameters θ^t, and $q(\theta^t)$ is the posterior distribution of model parameters, which cannot be obtained directly but can be estimated by dropout variational inference. Dropout variational inference is a practical method for approximating large and complex models [31]. Dropout is applied to every weight layer both in the training and the testing phase. The inference is obtained by sampling the approximate posterior, also referred to as Monte Carlo dropout. In addition, Liu et al. [32] believe that a prediction with high certainty should be consistent under randomly sampled subnetworks and random noise in inputs. Assuming that the set of sub-sampling results of K random enhancements of input data is $Y = \left\{\tilde{y}_{i,k}^t(x_i, \theta_k^t, \eta'_k)\right\}_{k=1}^K$, and the prediction variance (PV) [30] is used to measure the uncertainty of the prediction. The higher the variance, the higher the uncertainty:

$$\begin{aligned} \text{U}(x_i) = PV &= \sum_c Var[p(\tilde{y}_{i,1}^t = c | x_i, \theta_1^t, \eta'_1), \cdots, p(\tilde{y}_{i,K}^t = c | x_i, \theta_K^t, \eta'_K)] \\ &= \sum_c \left(\frac{1}{K}\sum_k (p(\tilde{y}_{i,k}^t = c | x_i, \theta_k^t, \eta'_k) - \mu_c)^2\right) \\ & where\ \mu_c = \frac{1}{K}\sum_k p(\tilde{y}_{i,k}^t = c | x_i, \theta_k^t, \eta'_k) \end{aligned} \tag{7}$$

For input data $x_i, i \in \{1, \cdots, B\}$ of the mini-batch, the uncertainty values of teacher predictions are $[U(x_1), \cdots, U(x_B)]$. The inputs are sorted in ascending order of the uncertainty to form the ordered input set $\{P_1, \ldots, P_B\}$. The reliable input samples set $M = \{P_1, \ldots, P_m\}$ contains m lowest uncertainty samples chosen from the ordered input set and is used for the consistency constraints. $m = \min(\beta e, B)$, e as the epoch, and β as the ramp-up coefficient. The number of samples selected by certainty will increase over time, as the teacher predictions will become more accurate during training.
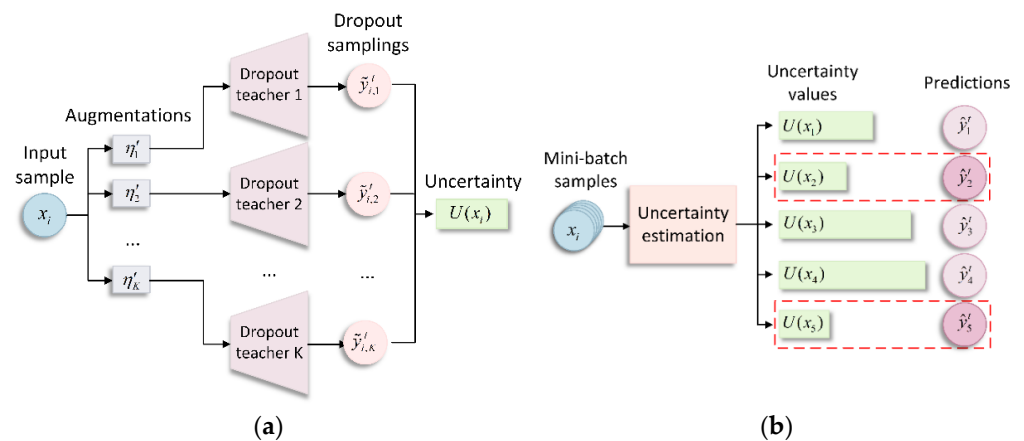
Figure 4. Certainty driven selection method: (**a**) uncertainty estimation; (**b**) predictions selection based on certainty.

2.3. Class-Imbalanced Learning

Imbalanced learning is a machine learning paradigm in which the classifier learns from dataset with a skewed class distribution. In this paper, we modify the training losses to further improve the robustness of the model to the imbalanced dataset. There are two main methods for solving class imbalance in supervised learning: loss re-weighting and mini-batch resampling [33–35]. These methods make the proportion of samples of different classes in the training loss closer to the test distribution to achieve a better trade-off between the accuracy of majority and minority classes. However, the model's scale is usually massive relative to the number of samples of the minority class, so there is the problem of over-fitting to the minority class. Label-distribution-aware margin loss (LDAM Loss) [28] regularizes different classes according to the number of samples: The regularization of the minority class should be stronger than that of the majority class to boost the generalization ability of the model to the minority class without sacrificing the fitting ability to the majority class. Figure 5 shows an example of binary classification, where χ_1 and χ_2 represent the margin of majority class and minority class, respectively. Class margin is the minimum distance from all samples of this class to the decision boundary. The minority class should have a more significant margin than the majority class. For the multi-classification problem, when the margin of class c is satisfied $\chi_c \propto 1/N_c^{1/4}$, the minimum test error can be obtained, where N_c is the sample number of class c.
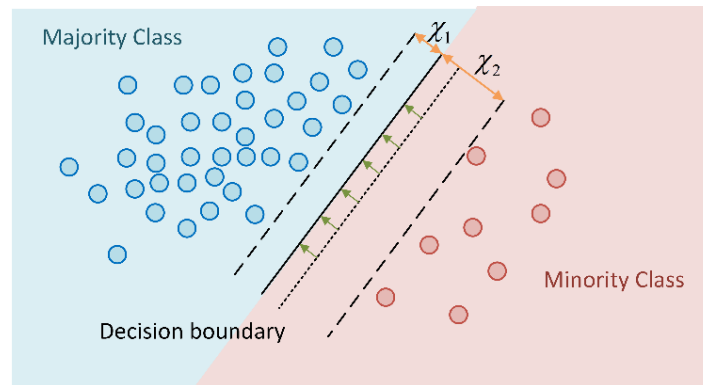
Figure 5. Adjusting class margin to reduce test errors.

Therefore, Hinge loss is adopted to enforce the class margin. For the labeled sample (x_i, y_i), hinge loss is

$$L_{hinge} = \max(\max_{l \neq y_i}\{z_l\} - z_{y_i} + \Delta_{y_i}, 0), \tag{8}$$

where z_l is the lth output of $\hat{y}_i^s$ predicted by the student model, Δ_{y_i} is the margin of class y_i, satisfying $\Delta_{y_i} = A/N_{y_i}^{1/4}$, and A is a constant for margin tuning. Since the hinge loss is non-convex and non-continuous, it is hard to optimize. A smoother cross-entropy loss with enforced class margin is adopted

$$L_{\text{LDAM}}(\hat{y}_i^s, y_i) = -\log \frac{e^{z_{y_i} - \Delta_{y_i}}}{e^{z_{y_i} - \Delta_{y_i}} + \sum_{l \neq y_i} e^{z_l}}, \tag{9}$$

The class imbalance will bring more challenges to the semi-supervised learning algorithm because, based on the smooth hypothesis, the decision boundary is located in the low-density area of the data space [23]. However, in the case of the class-imbalanced dataset, the high-density area of the minority class is sparse relative to the majority class, which causes that the decision boundary enters the minority class region, making the model predict the minority class samples to be the majority class. Therefore, to prevent the decision boundary from being overly smooth and infiltrating into the minority class areas, when the prediction given by the teacher model is the majority class, the consistency constraints should be suppressed. For m samples x_j from the certainty-driven selection, the compression consistency loss is defined as

$$L_{CCL}(\hat{y}_j^s, \hat{y}_j^t) = g(N_{\hat{c}})\left\|\hat{y}_j^s - \hat{y}_j^t\right\|^2$$
$$where \; g(N_{\hat{c}}) = \delta^{1 - \frac{N_{min}}{N_{\hat{c}}}} \tag{10}$$

$\hat{c}$ represents the class predicted by the model, $\delta \in (0, 1]$ is the compression coefficient, and $N_{\min}$ is the sample number of the class with the least samples. When $N_{\hat{c}} = N_{\min}$, $g(N_{\hat{c}}) = 1$, the compression consistency loss of samples predicted to be the smallest class is the same as the typical consistency regularization loss (CRL). The larger the data size of the predicted class $\hat{c}$, the smaller $g(N_{\hat{c}})$ is. Therefore, the final semi-supervised cost function is

$$J = -\sum_{i=1}^{b_l} \log \frac{e^{z_{y_i} - \Delta_{y_i}}}{e^{z_{y_i} - \Delta_{y_i}} + \sum_{l \neq y_i} e^{z_l}} + \sum_{x_j \in M} g(N_{\hat{c}})\left\|\hat{y}_j^s - \hat{y}_j^t\right\|^2 \tag{11}$$

3. Results

3.1. Dataset

The original images of the assembly were captured from the production line by the industrial camera BM-500GE (produced by JAI) with a CCD resolution of 2456 × 2058. We clipped the original images to form an assembly image dataset to reduce the influence of irrelevant background objects and focus on the central area of the assembly. The minimum resolution of images in the dataset was 279 × 235, and the maximum resolution was 313 × 528. All assembly images were divided into three classes: two unqualified classes (missing lead seal and missing lead seals and gaskets) and a qualified class. Three types of images in the dataset are shown in Figure 6. There are only fine-grained differences between the two unqualified classes, so distinguishing the two unqualified minority classes is a great difficulty.
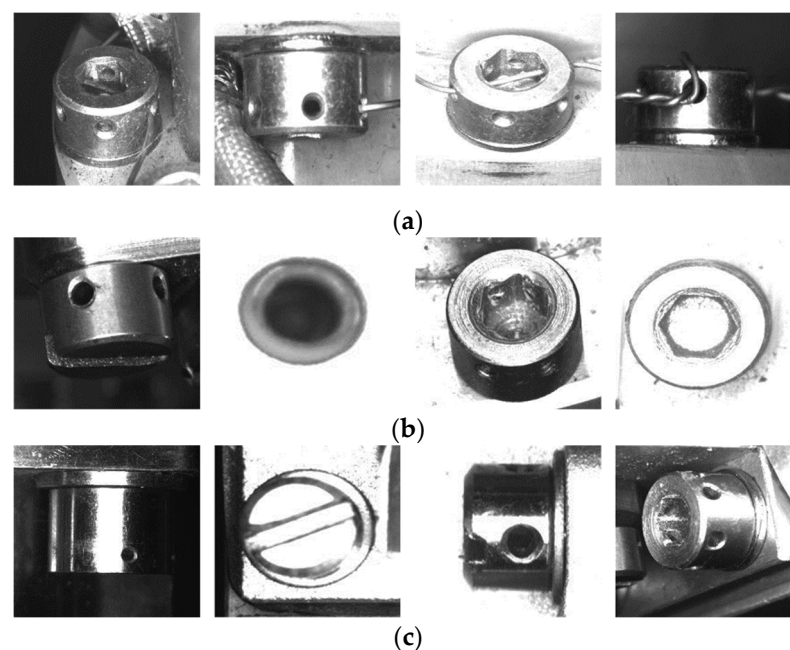

Figure 6. Examples of three classes of the dataset: (**a**) qualified; (**b**) missing lead seal and gasket; (**c**) missing lead seal.

The training set contained a total of 16,663 images. Three labeled fractions ε of 10%, 20%, and 50% were adopted in experiments. We randomly extracted images for manual labeling to ensure that the labeled and unlabeled datasets had the same distribution. The sample numbers of three classes in the training set and test set are shown in Table 1. It is worth noting that, to verify the model's classification performance on minority classes, we used a class-balanced test set in this study. In the labeled training set, the imbalanced ratio of the majority class to the minority class was as high as 5.3, which was severe for traditional classification methods.

Table 1. Image numbers of the classes in training set and testing set.

Class	Training Set			Testing Set
	$\varepsilon = 10\%$	$\varepsilon = 20\%$	$\varepsilon = 50\%$	
Qualified	1136	2272	5680	100
Both missing	311	622	1555	100
Sealing missing	216	432	1080	100
Unlabeled	15,000	13,337	8348	-

3.2. Training Settings and Metrics

The teacher and student were initialized by the pre-trained weights of DenseNet121 on ImageNet [36], and 100 epochs were performed with the mini-batch size of 16, where the batch size of labeled data was 8. Because the number of unlabeled images was no less than the labeled images, in every epoch, labeled images were iterated unlimited times until every unlabeled image was iterated once. We used stochastic gradient descent (SGD) to optimize the network, with a learning rate of 0.1, weight decay of 0.0001, and momentum of 0.9. The maximum value of EMA coefficient μ_0 of the teacher model was 0.999. To obtain the certainty values of teacher model predictions, we used MC dropout five times, and the ramp-up coefficient β was 2. For the LDAM, parameter A was tuned so that the maximum margin was 0.5. For CCL, δ was set to 0.5. Random augmentations applied to training data included random horizontal flipping and color jitter of brightness and contrast. After random augmentations, all images were resized to 224×224 before being sent to the network.

We used two common metrics in class-imbalanced learning to evaluate the performance of the model on the test set: balanced accuracy (bACC) [37] and geometric mean score (GM) [38], which are arithmetic and geometric mean scores, respectively, defined as

$$\text{bACC} = \frac{1}{C}\sum_{c=1}^{C}\frac{TP_c}{N_c},\tag{12}$$

$$\text{G} - \text{Mean} = \prod_{c=1}^{C}\sqrt[\frac{1}{C}]{\frac{TP_c}{N_c}},\tag{13}$$

where N_c represents the sample number of class c, and TP_c represents the sample number both belonging to class c and predicted to be class c.

We trained the proposed and conventional methods for comparison on the training sets with the labeled fraction ε of 10%, 20%, and 50%. Then the classification performances were compared on the test dataset. Conventional methods included: supervised learning with limited labeled data; The standard mean teacher method; The mean teacher method with three commonly used class-imbalanced learning strategies: (1) re-weighting: the loss of each sample was re-weighted by the inverse of the sample number of the corresponding class, and re-normalized to make the average weight in the mini-batch was 1; (2) resampling: the sampling probability of each sample was inversely proportional to the sample number in its class; (3) focal loss [39]: the loss of the relatively correctly classified sample was reduced, and the loss of the difficult and incorrectly classified sample was increased. To ensure fair comparisons, all methods adopted the DenseNet121 model structure and the same hyper-parameters as the proposed method, such as pre-training initialization, labeled data batch size, and the optimization method mentioned above. All the training and testing experiments were repeated ten times, and the experimental results on the test set were averaged. The algorithm in this paper was implemented using Python toolkit PyTorch, and experiments were carried out on a computer with Intel Core I5-8500 @ 3.00 GHz CPU and 12G NVIDIA Titan RTX GPU.

3.3. Experimental Results

Figure 7 shows the differences in the representation of the training set between the supervised learning method (a) and the proposed method (b). T-SNE [40] projection with perplexity of 50 was used for visualization. In Figure 7a, the boundaries of the three classes were mixed. Therefore, under the condition of limited and imbalanced data, the model trained by the conventional supervised learning method was hard to learn discriminative data representation. The proposed method could form better class boundaries and obtain better classification performance.
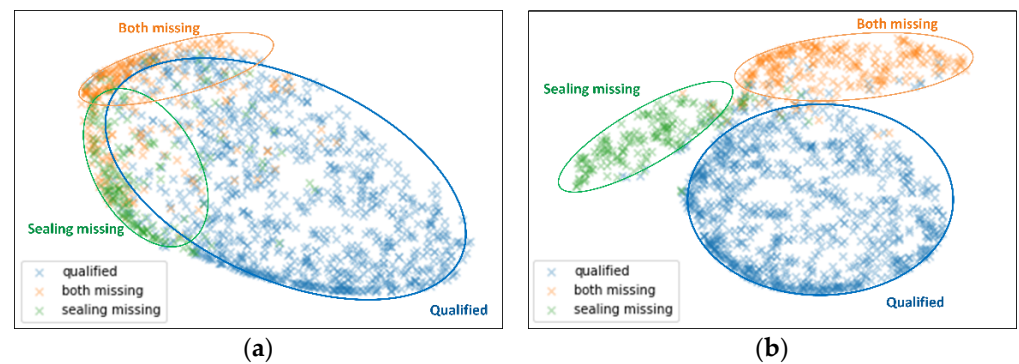
Figure 7. T-SNE visualization of the training set: (**a**) supervised learning; (**b**) the proposed method.

Table 2 shows the mean and standard deviation of bACC and GM for the above methods on the test set. The proposed algorithm achieved an average bACC of 93.67% and an average GM of 93.57% when the labeled fraction was 10%. When the labeled fraction was 20%, an average ACC of 98.83% and an average GM of 98.83% were achieved. When the labeled fraction was 50%, an average ACC of 99.17%, and an average GM of 98.99% were reached. The proposed method performed better than the supervised learning method and all the mean teacher methods with existing class-imbalanced learning strategies, indicating that the proposed method was effective in the case of limited labeled data with the imbalanced class distribution. In addition, with less annotated data, the proposed method had more advantages and had a higher accuracy than other methods.

Table 2. Classification performance comparison on the assembly test set (bACC(%)/GM(%)).

Methods	$\varepsilon = 10\%$		$\varepsilon = 20\%$		$\varepsilon = 50\%$	
	bACC	GM	bACC	GM	bACC	GM
Supervised	85.34 ± 0.94	85.10 ± 0.99	94.50 ± 0.42	94.32 ± 0.45	97.00 ± 0.09	96.98 ± 0.09
MT	88.22 ± 0.45	87.87 ± 0.47	95.50 ± 0.52	95.37 ± 0.54	98.16 ± 0.05	98.16 ± 0.05
MT + Reweight	89.34 ± 0.66	88.98 ± 0.70	94.67 ± 0.56	94.50 ± 0.61	97.84 ± 0.05	97.82 ± 0.04
MT + Resample	90.34 ± 0.47	90.06 ± 0.29	95.84 ± 0.52	95.80 ± 0.53	97.84 ± 0.05	97.82 ± 0.05
MT + Focal	88.50 ± 0.81	87.86 ± 0.90	93.33 ± 0.28	92.95 ± 0.32	97.50 ± 0.09	97.46 ± 0.11
Proposed method	93.67 ± 0.27	93.57 ± 0.28	98.83 ± 0.14	98.83 ± 0.14	99.17 ± 0.07	98.99 ± 0.04

Figure 8 shows the error rates of all methods in the three classes. Figure 8a–c resulted from training the models under the labeled fraction ε of 10%, 20%, and 50%, respectively. The proposed method kept a low error rate in the majority and minority classes. When labeled data were few (Figure 8a,b), the supervised learning method showed high error rates in all classes. In contrast, although the mean teacher method achieved higher accuracy, its error rates in the minority classes did not decrease significantly because of confirmation bias. The mean teacher method combined with class-imbalanced learning strategies led to overfitting in the lead seal missing class, which was the greatest minority. Although it achieved a lower error rate, it sacrificed its fitting ability in the similar sub-minority class–both lead seal and gaskets missing. In addition, with the gradual increase of labeled data, the supervised learning algorithm had already obtained a low error rate, and the mean teacher method with class-imbalanced learning strategies did not observably improve the error rate.

Figure 9 compares the accuracy and loss of the proposed method and the standard mean teacher method in the training process. The accuracy values were the validation accuracy after each epoch. A thousand uniform sampled loss values from all iterations were used to plot Figure 9(a2–c2). It can be seen that the proposed method converged faster and tended to be more stable, and a model with higher accuracy could be obtained with fewer iteration steps. In Figure 10, the accuracy and prediction variance (PV) during the training

process are compared. With the ascending of model accuracy, the uncertainty of prediction labels gradually decreased. There was a strong inverse relationship between classification accuracy rate and average PV, which verified certainty driven selection's effectiveness.
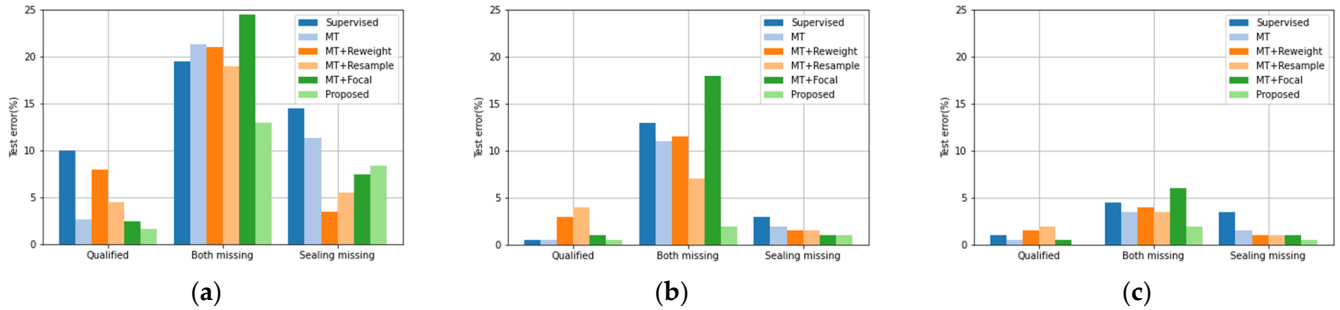

Figure 8. Test error rates from experiments under three percentages of labeled data: (**a**) $\varepsilon = 10\%$; (**b**) $\varepsilon = 20\%$; (**c**) $\varepsilon = 50\%$.
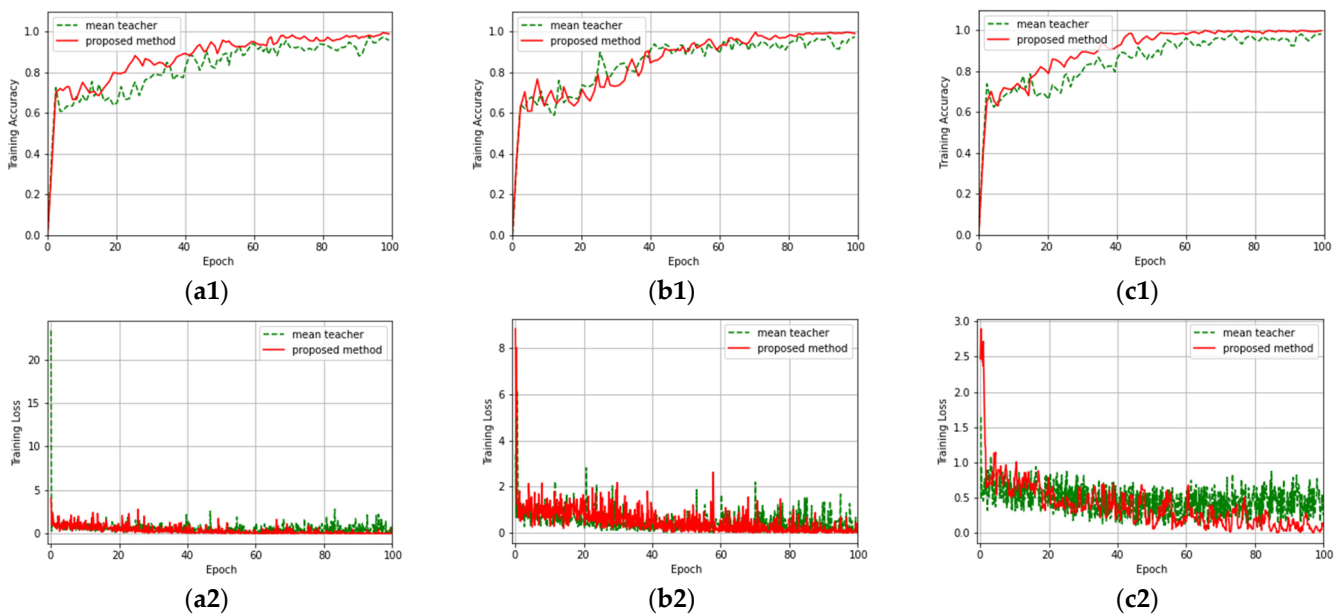

Figure 9. Training accuracy (top row) and loss (bottom row) comparison of mean teacher and the proposed method under three percentages of labeled data: (**a1,a2**) $\varepsilon = 10\%$; (**b1,b2**) $\varepsilon = 20\%$; (**c1,c2**) $\varepsilon = 50\%$.
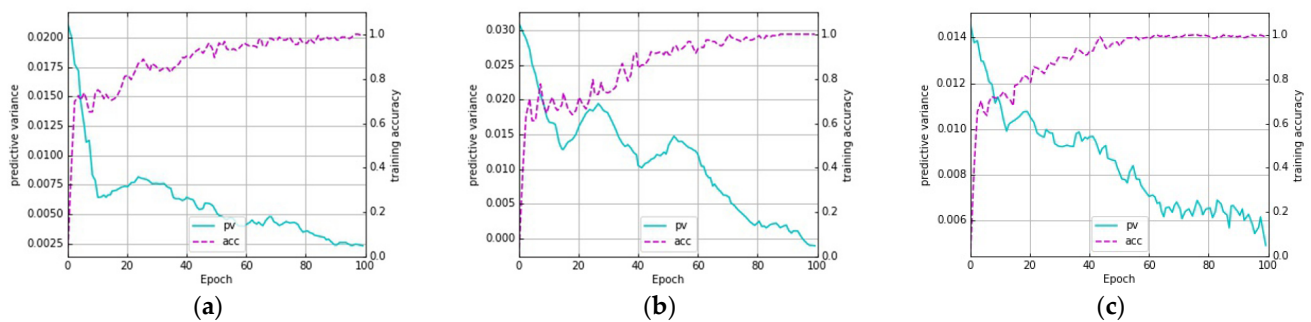

Figure 10. There was a strong inverse relationship between the accuracy and prediction variance during training: (**a**) $\varepsilon = 10\%$; (**b**) $\varepsilon = 20\%$; (**c**) $\varepsilon = 50\%$.

4. Conclusions

This paper represents a semi-supervised class-imbalanced learning method based on the mean teacher to detect unqualified assembly samples. For consistency constraints, samples with high reliability are selected according to the model prediction certainty to improve the performance. Label distributed aware margin loss and compression consistency loss are employed to guarantee the accuracy of classification without sacrificing the fitting ability to the majority class. Experiments were carried out on the assembly image dataset, and the performance was evaluated and compared with traditional deep learning classification methods. To verify the performance of the proposed method on a small amount of labeled data, 10% of the total data were labeled. Experimental results show that the prediction accuracies of the supervised learning method, mean teacher algorithm, and the proposed method were 85.34%, 88.22%, and 93.67% respectively. The proposed method overcame the performance degradation of the traditional semi-supervised learning algorithm on class-imbalanced datasets, kept low error rates in all classes, and could effectively avoid over-fitting on the minority class that occurred in the commonly used class-imbalanced learning methods. The model performances were also discussed when the labeled fraction increased to 20% and 50%, and the proposed method still achieved the highest accuracies of 98.83% and 98.99%. Future work will focus on applying the proposed method to more manufacturing scenarios and further enhancing classification accuracy.

References

1. Feng, H.; Jiang, Z.; Xie, F.; Yang, P.; Shi, J.; Chen, L. Automatic Fastener Classification and Defect Detection in Vision-Based Railway Inspection Systems. *IEEE Trans. Instrum. Meas.* **2014**, *63*, 877–888. [CrossRef]
2. Liu, L.; Zhou, F.; He, Y. Automated Visual Inspection System for Bogie Block Key Under Complex Freight Train Environment. *IEEE Trans. Instrum. Meas.* **2016**, *65*, 2–14. [CrossRef]
3. Wojciechowski, J.; Suszynski, M. Optical Scanner Assisted Robotic Assembly. *Assem. Autom.* **2017**, *37*, 434–441. [CrossRef]
4. Suszynski, M.; Wojciechowski, J.; Zurek, J. No Clamp Robotic Assembly with Use of Point Cloud Data from Low-Cost Triangulation Scanner. *Teh. Vjesn. Tech. Gaz.* **2018**, *25*, 904–909. [CrossRef]
5. Reyno, T.; Marsden, C.; Wowk, D. Surface Damage Evaluation of Honeycomb Sandwich Aircraft Panels Using 3D Scanning Technology. *NDT E Int.* **2018**, *97*, 11–19. [CrossRef]
6. Park, J.; Huynh, T.; Choi, S.; Kim, J. Vision-Based Technique for Bolt-Loosening Detection in Wind Turbine Tower. *Wind. Struct. Int. J.* **2015**, *21*, 709–726. [CrossRef]
7. Gibert, X.; Patel, V.M.; Chellappa, R. Deep Multitask Learning for Railway Track Inspection. *IEEE Trans. Intell. Transp. Syst.* **2017**, *18*, 153–164. [CrossRef]
8. Chen, J.; Liu, Z.; Wang, H.; Nunez, A.; Han, Z. Automatic Defect Detection of Fasteners on the Catenary Support Device Using Deep Convolutional Neural Network. *IEEE Trans. Instrum. Meas.* **2018**, *67*, 257–269. [CrossRef]
9. Kang, G.; Gao, S.; Yu, L.; Zhang, D. Deep Architecture for High-Speed Railway Insulator Surface Defect Detection: Denoising Autoencoder With Multitask Learning. *IEEE Trans. Instrum. Meas.* **2019**, *68*, 2679–2690. [CrossRef]
10. Chang, L.; Liu, Z.; Shen, Y.; Zhang, G. Novel Multistate Fault Diagnosis and Location Method for Key Components of High-Speed Trains. *IEEE Trans. Ind. Electron.* **2021**, *68*, 3537–3547. [CrossRef]

11. Wang, C.; Xu, Z. An Intelligent Fault Diagnosis Model Based on Deep Neural Network for Few-Shot Fault Diagnosis. *Neurocomputing* **2021**, *456*, 550–562. [CrossRef]
12. Zhang, S.; Ye, F.; Wang, B.; Habetler, T.G. Semi-Supervised Bearing Fault Diagnosis and Classification Using Variational Autoencoder-Based Deep Generative Models. *IEEE Sens. J.* **2021**, *21*, 6476–6486. [CrossRef]
13. Tang, L.; Xuan, J.; Shi, T.; Zhang, Q. EnvelopeNet: A Robust Convolutional Neural Network with Optimal Kernels for Intelligent Fault Diagnosis of Rolling Bearings. *Measurement* **2021**, *180*, 109563. [CrossRef]
14. Liu, C.; Wu, Y.; Liu, J.; Sun, Z.; Xu, H. Insulator Faults Detection in Aerial Images from High-Voltage Transmission Lines Based on Deep Learning Model. *Appl. Sci.* **2021**, *11*, 4647. [CrossRef]
15. Gao, Z.; Yang, G.; Li, E.; Liang, Z. Novel Feature Fusion Module-Based Detector for Small Insulator Defect Detection. *IEEE Sens. J.* **2021**, *21*, 16807–16814. [CrossRef]
16. Liu, J.; Liu, C.; Wu, Y.; Xu, H.; Sun, Z. An Improved Method Based on Deep Learning for Insulator Fault Detection in Diverse Aerial Images. *Energies* **2021**, *14*, 4365. [CrossRef]
17. Wang, Y.; Yao, Q.; Kwok, J.T.; Ni, L.M. Generalizing from a Few Examples: A Survey on Few-Shot Learning. *ACM Comput. Surv.* **2020**, *53*, 63:1–63:34. [CrossRef]
18. Wu, J.; Zhao, Z.; Sun, C.; Yan, R.; Chen, X. Few-Shot Transfer Learning for Intelligent Fault Diagnosis of Machine. *Measurement* **2020**, *166*, 108202. [CrossRef]
19. Gao, F.; Ma, F.; Wang, J.; Sun, J.; Yang, E.; Zhou, H. Semi-Supervised Generative Adversarial Nets with Multiple Generators for SAR Image Recognition. *Sensors* **2018**, *18*, 2706. [CrossRef]
20. Cabrera, D.; Sancho, F.; Long, J.; Sánchez, R.; Zhang, S.; Cerrada, M.; Li, C. Generative Adversarial Networks Selection Approach for Extremely Imbalanced Fault Diagnosis of Reciprocating Machinery. *IEEE Access* **2019**, *7*, 70643–70653. [CrossRef]
21. Li, X.; Zhang, W.; Ding, Q. Cross-Domain Fault Diagnosis of Rolling Element Bearings Using Deep Generative Neural Networks. *IEEE Trans. Ind. Electron.* **2019**, *66*, 5525–5534. [CrossRef]
22. Zhou, F.; Song, Y.; Liu, L.; Zheng, D. Automated Visual Inspection of Target Parts for Train Safety Based on Deep Learning. *IET Intell. Transp. Syst.* **2018**, *12*, 550–555. [CrossRef]
23. Chapelle, O.; Schölkopf, B.; Zien, A. *Semi-Supervised Learning*; MIT Press: Cambridge, MA, USA, 2009.
24. Sajjadi, M.; Javanmardi, M.; Tasdizen, T. Regularization With Stochastic Transformations and Perturbations for Deep Semi-Supervised Learning. In Proceedings of the 30th International Conference on Neural Information Processing Systems, Barcelona, Spain, 5–10 December 2016; Curran Associates Inc.: Red Hook, NY, USA, 2016.
25. Laine, S.; Aila, T. Temporal Ensembling for Semi-Supervised Learning. In Proceedings of the 5th International Conference on Learning Representations, ICLR 2017, Toulon, France, 24–26 December 2017.
26. Tarvainen, A.; Valpola, H. Mean Teachers Are Better Role Models: Weight-Averaged Consistency Targets Improve Semi-Supervised Deep Learning Results. In Proceedings of the 31st International Conference on Neural Information Processing Systems, Long Beach, CA, USA, 4–9 December 2017; Curran Associates Inc.: Red Hook, NY, USA, 2017.
27. Kim, J.; Hur, Y.; Park, S.; Yang, E.; Hwang, S.; Shin, J. Distribution Aligning Refinery of Pseudo-Label for Imbalanced Semi-Supervised Learning. In Proceedings of the 34th International Conference on Neural Information Processing Systems, Online, 6–12 December 2020; Curran Associates Inc.: Red Hook, NY, USA, 2020.
28. Cao, K.; Wei, C.; Gaidon, A.; Arechiga, N.; Ma, T. Learning Imbalanced Datasets with Label-Distribution-Aware Margin Loss. In Proceedings of the 33rd International Conference on Neural Information Processing Systems, Vancouver, BC, Canada, 8–14 December 2019; Curran Associates Inc.: Red Hook, NY, USA, 2019.
29. Huang, G.; Liu, Z.; Maaten, L.V.D.; Weinberger, K.Q. Densely Connected Convolutional Networks. In Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, USA, 21–26 July 2017; pp. 2261–2269.
30. Gal, Y.; Ghahramani, Z. Dropout as a Bayesian Approximation: Representing Model Uncertainty in Deep Learning. In Proceedings of the 33rd International Conference on Machine Learning, New York, NY, USA, 19–24 June 2016.
31. Gal, Y.; Ghahramani, Z. Bayesian Convolutional Neural Networks with Bernoulli Approximate Variational Inference. In Proceedings of the 4th International Conference on Learning Representations, ICLR2016. Caribe Hilton, San Juan, Puerto Rico, 2–4 May 2016.
32. Liu, L.; Li, Y.; Tan, R.T. Decoupled Certainty-Driven Consistency Loss for Semi-Supervised Learning. *arXiv* **2020**, arXiv:1901.05657.
33. He, H.; Garcia, E.A. Learning from Imbalanced Data. *Knowl. Data Eng. IEEE Trans.* **2008**, *21*, 1263–1284. [CrossRef]
34. Buda, M.; Maki, A.; Mazurowski, M.A. A Systematic Study of the Class Imbalance Problem in Convolutional Neural Networks. *Neural Netw.* **2018**, *106*, 249–259. [CrossRef]
35. Cui, Y.; Jia, M.; Lin, T.-Y.; Song, Y.; Belongie, S. Class-Balanced Loss Based on Effective Number of Samples. In Proceedings of the 2019 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Long Beach, CA, USA, 16–20 June 2019.
36. Deng, J.; Dong, W.; Socher, R.; Li, L.; Li, K.; Li, F.-F. ImageNet: A Large-Scale Hierarchical Image Database. In Proceedings of the 2009 IEEE Conference on Computer Vision and Pattern Recognition, Miami, FL, USA, 20 June 2009; pp. 248–255.
37. Huang, C.; Li, Y.; Loy, C.C.; Tang, X. Learning Deep Representation for Imbalanced Classification. In Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, USA, 26 June–1 July 2016; pp. 5375–5384.
38. Branco, P.; Torgo, L.; Ribeiro, R.P. A Survey of Predictive Modeling on Imbalanced Domains. *ACM Comput. Surv.* **2016**, *49*, 1–50. [CrossRef]

39.   Lin, T.-Y.; Goyal, P.; Girshick, R.; He, K.; Dollár, P. Focal Loss for Dense Object Detection. In Proceedings of the 2017 IEEE International Conference on Computer Vision, Venice, Italy, 22–19 October 2017; pp. 2980–2988.

40.   Van der Maaten, L.; Hinton, G. Visualizing Data Using T-SNE. *J. Mach. Learn. Res.* **2008**, *9*, 2579–2605.