Proceedings of the 4th International Conference on Mechatronics and Smart Systems
DOI: 10.54254/2755-2721/2026.34214

Adaptation Strategies for Few-shot Industrial Defect Image Classification

Zitian Li

School of Information Science and Engineering, Chongqing Jiaotong University, Chongqing, China
lizitian-cn@outlook.com

Abstract. In the current industrial era, artificial intelligence is gradually integrated into various industries, bringing more possibilities to production. In industrial manufacturing, surface defect detection is a key part of product quality control, but there are still problems of unsatisfactory efficiency. Therefore, this paper compares the performance of five adaptation strategies, including Linear Probing (LP) and Full Fine-tuning (FT), in small-sample industrial defect detection through the transfer learning framework based on ImageNet pre-trained ResNet18, aiming to provide method guidance for small-sample industrial defect detection under different sample data volumes. On the NEU-DET dataset, the prototype network is the best when the sample size is extremely small, and the accuracy of full fine-tuning reaches the highest 89.28% when the sample size is 5-shot and above, and the low-rank adaptive method only achieves 96.4% of the full fine-tuning performance with only 0.73% of the trainable parameters. DAGM 2007 Class 10 cross-dataset validation shows that full fine-tuning still maintains its advantage.

Keywords: industrial defect detection, small sample learning, transfer learning, low-rank adaptation, full fine-tuning

1. Introduction

Deep learning-based detection methods have gradually replaced traditional manual product detection methods due to their fast detection speed and high accuracy. However, in the actual production process, the detection results of this method are often unsatisfactory due to factors such as the scarcity of sample data that cannot be trained to the optimal neural network, and the surface of most industrial components has small defects that are highly similar to the inspection background [1, 2]. Texture plays an important role in image detection. It appears in almost all images, and because of its indescribable features, the traditional deep learning method requires a large amount of annotated data for training to achieve sufficient accuracy. This method is expensive and time-consuming [3].

Transfer learning is a new machine learning method that uses existing knowledge to solve problems in different but related fields. Compared with traditional machine learning, this method does not need enough available training samples, but can transfer existing knowledge to the target field with a small number of labeled samples or even no labeled samples, so as to solve the problem [4].

DOI: 10.54254/2755-2721/2026.34214

By improving the low rank adaptive (LORA) method in the parameter-efficient fine-tuning method, Huang proposed the FFLOR method, which significantly improved the performance of Sam in the image segmentation task of steel surface defects, and only needed to fine-tune the parameters of the model by 3.2% [5]. Zeng conducted full parameter fine-tuning training on yolov11 (you only look once version 11) through transfer learning to improve the generalization ability of the model and provide technical support for the safe operation of small hydropower station equipment [6]. The demand for defect detection under the condition that there are only a small number of label samples or even no in the target field is gradually increasing. The research of using the pre training model of large-scale data sets such as ImageNet to learn to the target field through migration is more important. However, in the neighborhood of industrial defect detection, due to the variability of its conditions and the serious scarcity of data, there is still a lack of systematic research on how to choose appropriate adaptation strategies to reduce the cost of industrial production. According to the parameter update method, the adaptation strategy is divided into feature extraction represented by linear detection. By freezing the pre training backbone network and only training the classification header of new tasks, the efficient calculation is realized, but the adaptability is limited [7]. Full fine tuning can train all parameters end-to-end and has the strongest expression ability, but it is easy to over fit in small sample scenes [8]. Parameter efficient fine tuning (PEFT) is a way to update only a few parameters and keep most of them frozen. Lora, which injects a trainable matrix through low rank decomposition, has achieved success in the field of natural language processing and has also been used to deal with computer vision problems in recent years [9].

This study provides a scientific suggestion for the selection of industrial defect detection methods by comparing five migration strategies.

2. Method

2.1. Data set

This paper selects the Neu-Det data set released by Northeast University, which includes six types of common defects. The data set has high annotation quality and can effectively evaluate the performance of the model in defect detection and classification. At the same time, the data set meets the small sample scenario, and the high similarity between different categories adds challenges to this study [10]. In addition, the DAGM 2007 class 10 dataset, which is closer to the industrial application scenario, is used in this study to cover strong supervision and weak supervision, avoiding the contingency caused by a single dataset [11].

2.2. Experimental pretreatment

ResNet-18is selected as the backbone network in this experiment, and the total training parameter is $11.7M$. The input size is 224×224 . For the RGB-based NEU-DET, the network retains a convolution layer with 3 input channels and 64 output channels in the first layer. For gray-scale DAGM, the corresponding modification of the first layer convolution is that the number of input channels is 1, the number of output channels is 64, and the size, step size, and filling of the other convolution cores remain unchanged at 7, 2, and 3, respectively.

DOI: 10.54254/2755-2721/2026.34214

2.3. Comparative experiment

In this experiment, 42, 123, 456 three groups of random seeds were used to eliminate the random effect.

LAs the most conservative migration strategy, LP can quickly verify the quality of pre training features under extremely limited resources by freezing the pre training model f_0 and training only the top linear classifier. Quality of training characteristics. On neu-det dataset, the trainable parameter of this method is 3078, while on DAGM dataset, it is 1026. $AdamW$ optimizer is used for model training, and its super parameter setting learning rate is $1*10^{-4}$, and the weight attenuation coefficient is $1*10^{-2}$.

The training linear classifier is

$$\hat{y} = softmax(W \bullet f_0(x) + b) \min_{W,b} \frac{1}{N} \sum_{i=1}^{N} L_{\mathrm{CE}}\left(y_i, \hat{y}_i\right) \tag{1}$$

$$W \in R^{C*d}, b \in R^{C} \tag{2}$$

Where C is the number of categories and d is the feature dimension

FT trains end-to-end by unfreezing all 11.2M trainable parameters. The optimizer also configures a learning rate of $1*10^{-4}$, which has a significant effect when labeling data and computing resources are sufficient.

$$\theta^{*} = arg\min_{\theta} \frac{1}{N} \sum_{i-1}^{N} L\left(f_0(x_i), y_i\right) + \lambda ||\theta||^2 \tag{3}$$

Progressive unfreezing (PU) is a step-by-step unfreezing method. First, fine-tune the high level and then gradually add the low level. In the first stage, $Epoch1-10$, only the classification head is trained, and the learning rate is $1*10^{-4}$. $Epoch11-20$ unfrozen the whole network, and the learning rate dropped to $1*10^{-5}$.

Lora (low rank adaptation) is to freeze the original model and add a low rank trainable matrix in the key layer to achieve a better model under the condition of limited training parameters.

$$W^{'} = W_0 + \Delta W = W_0 + BA,\ h = W_0 x + B(Ax) \tag{4}$$

$$W_0 \in R^{d*k}, B \in R^{d*r}, A \in R^{r*k} r \ll min(d,k), InitializeB = 0 \tag{5}$$

Among them, rank $r = 16$, input dimension $d = 16$, output dimension $k = 512$, and the regularization strategy with dropout rate of 0.1 is adopted.

Prototypical networks (ProtoNet) have significant advantages in small sample conditions, without parameter update. In addition, the $AdamW$optimizer is used in this experiment. The first-order estimated decay rate $\beta_1 = 0.9$, the second−order estimated decay rate $\beta_2 = 0.999$, and the numerical stability term $\varepsilon = 1*10^{-8}$ cross entropy loss $L_{CE} = -\sum_{i=1}^{N} y_i \log\left(\hat{y}_i\right)$, where y_i , $\hat{y}_i$ respectively represent the real label and model prediction probability of the i th sample $Batch\ size = 8$, $training = 20$ rounds.

DOI: 10.54254/2755-2721/2026.34214

3. Results

3.1. NEU-DET dataset experimental results

Table 1. NEU-DET comparison table of the accuracy rate of each shot and each strategy

Method	1-shot	5-shot	10-shot	20-shot
LP	23.46 ± 4.05	41.75 ± 7.11	57.41 ± 2.07	75.64 ± 1.65
FT	46.72 ± 8.31	83.58 ± 1.94	83.27 ± 0.87	89.28 ± 3.09
PU	29.82 ± 5.14	74.18 ± 1.63	80.69 ± 1.35	87.39 ± 0.93
LoRA	24.98 ± 3.79	68.20 ± 3.93	76.41 ± 1.31	86.11 ± 1.33
ProtoNet	59.90	78.32	81.60	83.61

As shown in Table 1, the experimental results reveal a significant strategy selection effect:

Under the condition of $1-shot$, the accuracy of protonet reached 59.90%, which was significantly higher than the other four adjustment strategies. According to the nature of the prototype network, this method can effectively alleviate the over fitting problem through class center aggregation, so this method will be significantly better than the other four adjustment methods under the condition of extremely small samples of $1-shot$. When $5-shot$and above, the highest accuracy is achieved in FT mode. It is not difficult to find from Table 1 that when the sample size gradually increases, the advantage of FT gradually decreases. Among them, the training accuracy of Su is higher when the sample size gradually increases beyond 5-shot, and it is only 1.89%lower than that of FT method at $20-shot$, but under the limit of $1-shot$, the accuracy is far lower than that of protonet and ft. In addition, the Lora method, which trained 81920parameters only $0.73\%$of FT, achieved 86.11% accuracy at $20-shot$ and 96.4% accuracy of ft.

3.2. DAGM 2007 class10 cross dataset validation

Table 2. DAGM comparison table of accuracy rate of each shot and each strategy

Method	10-shot	20-shot	50-shot	100-shot
LP	51.60 ± 11.37	48.49 ± 8.50	48.66 ± 4.31	50.09 ± 7.92
FT	72.06 ± 12.16	54.32 ± 5.95	55.62 ± 20.16	74.55 ± 2.47
PU	62.49 ± 15.61	50.58 ± 4.22	53.16 ± 2.72	67.74 ± 3.06

DOI: 10.54254/2755-2721/2026.34214

Table 2. (continued)

LoRA	48.89 ± 21.64	52.61 ± 8.99	53.22 ± 6.56	55.91 ± 12.80
ProtoNet	50.17	53.39	52.61	53.48

As shown in Table 2, on the DAGM 2007 class10, the FT strategy achieved an accuracy of $72.06\% \pm 12.16\%$ and $74.55\% \pm 2.47\%$ under the conditions of $10-shot$ and $100-shot$, respectively, and the overall performance was the best, which proved the effectiveness of the strategy in the weak supervision scenario. However, the performance fluctuates at $20-shot$ and $50-shot$, hich is speculated to be the background noise introduced by image level labels and the risk of over-fitting under the condition of small samples. Compared with ft, the Lora method performs poorly on this dataset. It is speculated that the reason is that the texture defect area of DAGM is small and the contrast is low, while Lora only fine-tunes the last convolution layer, which makes it difficult to fully capture the multi-scale defect features. In addition, the standard deviation of all strategies is mostly higher than that of the NEU-DET data set, which reflects the inherent training instability of weakly supervised tasks. The accuracy ranking of the five strategies under this data set is basically consistent with that under the NEU-DET data set, which verifies the universality of the experimental conclusion.

3.3. Visualization of the strategy change phenomenon

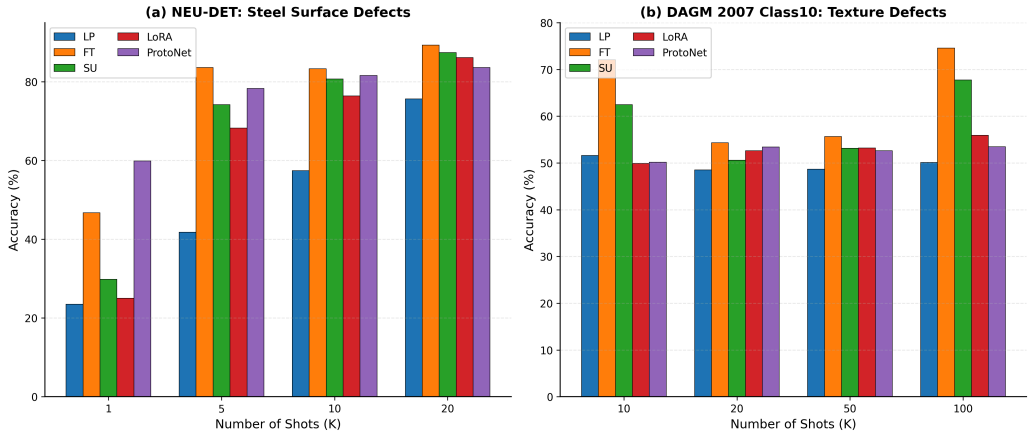

Figure 1. Strategy comparison histogram (photo/picture credit: original)

Figure 1 (a) visually shows that 1-shot protonet is in the lead under neu-det data set, and FT remains in the lead after 5-shot, Figure 1 (b) it shows that FT is always in the lead but fluctuates significantly under DAGM data set, and the accuracy of all adaptation strategies is generally lower than that under NEU-DET data set under the same conditions, reflecting the high challenge under weak supervision environment.

DOI: 10.54254/2755-2721/2026.34214

3.4. Parameter efficiency analysis

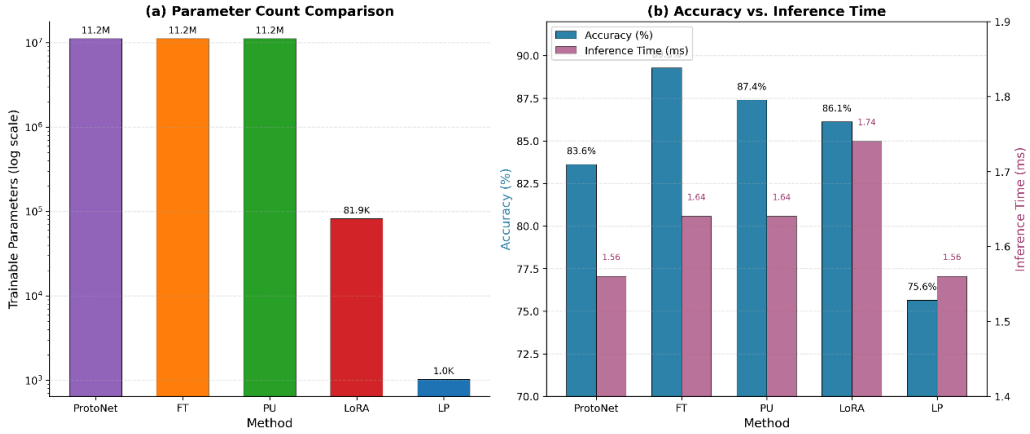

Figure 2. Comparison diagram of parameter efficiency (photo/picture credit: original)

In industrial production, the larger the number of training parameters, the greater the amount of computation (FLOPs) and memory consumption. Figure 2 (a) shows different adaptation strategies on the horizontal axis and training parameters on the vertical axis. It can be seen from Figure 2 that the training parameters of the linear detection method are far less than all adaptation strategies, followed by Lora. The horizontal axis of Figure 2(b) represents different adaptation strategies, and the vertical axis represents the accuracy and reasoning time under the $20-shot$ condition of the neu-det dataset. It can be seen from Figure 2 that the reasoning time of LP and protonet is lower than that of the other three adaptive strategies, which can greatly speed up the rate of industrial production, but at the same time, due to their poor accuracy, they are not suitable for high-precision demand scenarios.

3.5. Deficiencies and improvements

Although this study systematically compares the performance of five different adaptive strategies in small sample industrial defect classification, there are still limitations. At the method level, this experiment only studies a single adaptive strategy and does not try the mixed model, which limits the further extension of the universality of the conclusion. In terms of experimental design, NEU-DET and DAGM data sets are relatively small, and only accuracy, reasoning time, and training parameters are used as evaluation criteria, without considering the impact factors such as memory occupation in industrial deployment. Future improvements should be combined with the current cutting-edge visual detection methods to develop an adaptive adaptation framework for task perception, and automatically select the direction of the optimal strategy according to the sample size and task complexity.

4. Conclusion

This study evaluated the performance of five transfer learning adaptation strategies in small sample industrial defect classification tasks through a cross-dataset experimental system, and revealed the internal correlation mechanism between sample size, task complexity, and strategy selection. The research shows that under the condition of extremely small samples, the prototype network based on metric learning achieves the optimal generalization performance by virtue of its structural advantage

DOI: 10.54254/2755-2721/2026.34214

of feature space, which verifies the effectiveness of the metric learning method without parameter update in the case of sample scarcity; With the increase of training sample size, the full tuning strategy gradually shows performance advantages through sufficient parameter optimization, but its sensitivity to label noise is particularly significant in the weak supervision scenario; The two-stage fine-tuning strategy achieves a balance between performance and stability through the gradual thawing mechanism, providing a reliable solution for medium sample size scenarios; The performance of the low rank adaptation method is close to full tuning with minimal parameter cost, which confirms the application potential of the parameter efficient tuning method under resource constraints. Comparative analysis across datasets shows that the task complexity and the quality of the surveillance signal jointly determine the boundary conditions of strategy selection: the monotonicity of performance growth with sample size in the strong surveillance scenario is broken in the weak surveillance scenario, indicating that the impact of label quality on model performance is no less than the number of samples. The core enlightenment of this discovery for industrial applications is that when constructing a defect classification system, priority should be given to the rationality of labeling granularity, rather than blindly pursuing the sample size. The strategy selection framework proposed in this study provides an operable decision-making basis for the application of industrial defect classification - the prototype network is preferred for extremely small sample scenarios, the full amount of fine-tuning is used for high-precision demand scenarios, the low rank adaptation is selected for resource constrained scenarios, and the two-stage fine-tuning is considered for weak supervision scenarios - its engineering value lies in transforming the abstract small sample learning theory into a specific technical selection guide, reducing the threshold for the application of advanced algorithms in industry.

References

[1] Zhang, X. (2025). Industrial defect detection method based on small sample learning (Master's thesis). Harbin Institute of Technology.
[2] Chen, Z., Feng, X., Liu, L., & Jia, Z. (2023). Surface defect detection of industrial components based on vision. Scientific Reports, 13(1), 22136.
[3] Liu, L., & Kuang, G. (2009). Overview of image texture feature extraction methods. Journal of Image and Graphics, 14(04), 622–635.
[4] Zhuang, F., Luo, P., He, Q., & Shi, Z. (2015). Research progress of transfer learning. Journal of Software, 26(1), 14–39.
[5] Huang, M., & Yang, J. (2025). Efficient fine tuning of visual large model parameters for steel surface defect image segmentation. Mechanical Manufacturing, 63(5).
[6] Zeng, T., Chen, J., Xie, J., Pan, X., & Zhang, W. (2025). Intelligent identification method of generator carbon brush ignition in small hydropower station based on YOLOv11 full parameter fine tuning. Small Hydropower, 47(1), 13–17.
[7] Chen, W. Y., Liu, Y. C., Kira, Z., Wang, Y. C. F., & Huang, J. B. (2019). A closer look at few-shot classification. Paper presented at the International Conference on Learning Representations (ICLR), New Orleans, LA.
[8] Yosinski, J., Clune, J., Bengio, Y., & Lipson, H. (2014). How transferable are features in deep neural networks? In Advances in Neural Information Processing Systems (Vol. 27, pp. 3320–3328). Curran Associates, Inc.
[9] Hu, E. J., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., Wang, L., & Chen, Weizhu. (2022). LoRA: Low-rank adaptation of large language models. Paper presented at the International Conference on Learning Representations (ICLR).
[10] Song, K., & Yan, Y. (2013). A noise robust method based on completed local binary patterns for hot-rolled steel strip surface defects. Applied Surface Science, 285, 858–864.
[11] DAGM. (2007). DAGM GCPR 2007 workshop dynamic vision and pattern recognition competition.