# Few-shot Industrial Defect Image Classification Based on Lightweight Model with Attention Mechanism

Meiqi, Tu
Department of Automation, Tsinghua University, Beijing, China
tmq21@mails.tsinghua.edu.cn

Zhixiao, Qi
Department of Automation, Tsinghua University, Beijing, China
qzx21@mails.tsinghua.edu.cn

Libin, Yu
Urumqi Power Supply Section, Urumqi Railway Bureau, Xinjiang Uygur Autonomous Region, China
13565849451@139.com

Linxuan, Zhang*
Department of Automation, Tsinghua University, Beijing, China
lxzhang@tsinghua.edu.cn

## ABSTRACT

This study investigates the application of deep learning methods in industrial defect image classification, particularly when training samples are limited. A metric-based approach is proposed, which utilizes a pre-trained deep convolutional neural network for feature extraction. This approach achieves effective category discrimination by computing the cosine similarity between query images and support images, without the need for additional adjustments for new defect categories. To enhance feature extraction, a DenseNet with an attention mechanism is employed, providing a more lightweight model compared to ResNet12. The inclusion of a hybrid domain attention mechanism improves performance and alleviates potential performance degradation that may arise from parameter reduction. Extensive evaluations are conducted on both general datasets and industrial defect datasets, demonstrating the effectiveness of the proposed model in real-world scenarios. This approach only requires a small number of defect samples, and the attention mechanism-enhanced DenseNet feature extraction network utilizes only one-fourth of the parameters of ResNet12, achieving comparable or better detection results. An ablation experiment confirms the superiority of the DenseNet with an attention mechanism. In summary, this research contributes to the field of few-shot learning in industrial defect image classification by proposing a low-cost and efficient solution.

## CCS CONCEPTS

• **Computing methodologies** → Artificial intelligence; Computer vision; Computer vision problems.

## KEYWORDS

Few-shot Learning, Image Classification, Densely Connected Convolutional Network, Attention Mechanism

This work is licensed under a Creative Commons Attribution International 4.0 License.

*ASSE 2023, October 27–29, 2023, Aizu-Wakamatsu City, Japan*
© 2023 Copyright held by the owner/author(s).
ACM ISBN 979-8-4007-0853-4/23/10
https://doi.org/10.1145/3634814.3634841

**ACM Reference Format:**
Meiqi, Tu, Zhixiao, Qi, Libin, Yu, and Linxuan, Zhang*. 2023. Few-shot Industrial Defect Image Classification Based on Lightweight Model with Attention Mechanism. In *2023 4th Asia Service Sciences and Software Engineering Conference (ASSE 2023), October 27–29, 2023, Aizu-Wakamatsu City, Japan.* ACM, New York, NY, USA, 8 pages. https://doi.org/10.1145/3634814.3634841

## 1 INTRODUCTION

With the advancement in data acquisition, storage, processing capabilities, and the development of computational techniques, the application of deep learning methods in image classification has gained considerable attention. LeNet [1], the earliest proposed convolutional neural network model, successfully achieved handwritten digit recognition in the MNIST dataset by utilizing the network structure composed of convolutional layers, pooling layers, and fully connected layers, which laid the foundation for the development of deep convolutional neural networks. In 2012, AlexNet [2] was introduced, which employed the rectified linear unit (ReLU) as the activation function, introduced local response normalization to alleviate the issue of gradient vanishing, and effectively mitigated overfitting by employing data augmentation and dropout techniques. In 2014, VGG16 and VGG19 [3] validated the superior performance of stacking multiple small-sized convolutional kernels (3x3) and max-pooling kernels (2x2) compared to using individual large-sized convolutional kernels. GoogLeNet [4] adopted the Inception module, which reduced the model parameter count while ensuring computational efficiency by utilizing sparse connections. This significantly improved the network's performance even with a depth of 22 layers. ResNet [5] consisted of stacked residual blocks, where skip layers integrated output information from the previous two layers as input to the current layer. This approach allowed gradient information from upper layers to directly enter deeper layers, addressing the issue of gradient vanishing caused by continuous derivative calculations. By stacking residual blocks, ResNet achieved a depth of 152 layers and achieved significant success in image classification tasks. Subsequently, most deep convolutional networks have been improved based on the ResNet framework.

The advantages of deep learning-based image classification methods lie in their ability to extract deeper image features without relying on manual feature design. As a result, they have been applied to various scenarios, including industrial defect image classification. However, in the industrial sector, it is common to encounter

difficulties in obtaining sufficient training images due to factors such as sample scarcity, lack of annotations, data confidentiality, inadequate management measures, and limited storage conditions. Furthermore, the scarcity of data resources often coexists with inadequate computational resources. Unfortunately, the current mainstream deep learning-based image classification methods still rely on data-driven approaches to ensure model accuracy and good generalization. Moreover, when the number of classification categories increases, it is often necessary to readjust the model to adapt to the new scenarios, which consumes a significant amount of time and computational resources. Therefore, it is crucial to find a method for automatic image classification that does not require excessive annotation of sample images and computational resources.

The emergence of methods for image classification with small sample sizes has provided a valuable research direction to address the aforementioned need. Humans possess the ability to learn quickly, being able to recognize a new object with only a small number of samples. The problem that Few-shot Learning aims to solve is how to enable machine learning models to learn quickly with a small number of samples for new categories, after having learned from a large amount of data for certain categories. The core of Few-shot Learning models lies in how to introduce prior knowledge through different methods to mitigate the risk of data scarcity [6]. Without altering the model, data augmentation techniques can be employed to increase the quantity of data and thus reduce estimation errors. However, increasing the amount of data is not always an easy task. FSL methods can be broadly categorized into three types: Model Based, Metric Based, and Optimization Based. Model Based methods update parameters quickly on a small sample set by designing model structures that directly establish a mapping function between inputs and prediction values. Metric Based methods classify by measuring the distance between samples in the query set and samples in the support set, employing the idea of nearest neighbors. Optimization Based methods believe that conventional gradient descent methods struggle to fit in few-shot scenarios, hence adjusting the optimization approach using prior knowledge to achieve small sample classification tasks. Currently, most methods employ meta-learning to address small sample tasks. However, learning classification models directly as a classification task within the training set yields better classification performance compared to task-based meta-learning methods. In order to explore the issue of feature discrepancy obtained from meta-learning, Chen et al. proposed the Meta-Baseline [7] for small sample image classification. The Meta-Baseline defines a network that is first pre-trained on a data-rich base dataset for image classification tasks, and then the encoder obtained from pre-training is used to extract image features. In the Classifier-Baseline, a metric-based approach is used for classifying new categories by computing the cosine similarity between images in the query set and images in the support set to determine the image classes. The Meta-Baseline introduces a parameter when calculating cosine similarity to train the classifier in a meta-learning manner for improved detection performance.

Due to cost constraints in practical scenarios, compressing model parameters to reduce memory consumption and make the model lightweight has been a major focus of research for engineers. Lightweight networks refer to networks with small parameter size and low computational complexity. Examples of lightweight networks include SqueezeNet [8], Xception [9], MobileNet series [10], and ShuffleNet series [11]. The SqueezeNet network consists of fire modules, which are composed of squeeze and expand convolutional layers. The former only contains 1x1 convolutions, while the latter contains both 1x1 and 3x3 convolutions. The MobileNetv1 network utilizes depth-wise separable convolution and introduces hyperparameters like width multiplier and resolution multiplier, allowing for different model sizes depending on the application. In addition to improving the convolutional approach, changing the model's connectivity is also an important way to reduce parameter size. DenseNet, known for its dense connections, achieves parameter reduction through feature reuse. DenseNet [12] is composed of dense blocks, where each layer gains additional input from all preceding layers and passes its own feature maps to all subsequent layers using the Concatenation method. Each layer incorporates collective knowledge from previous layers. However, the constant duplication of feature maps consumes memory. Therefore, this paper employs a memory-efficient version of the DenseNet feature extraction network, implemented by Pleiss et al [13].

In order to improve the performance of the feature extraction network, an attention mechanism is considered to be added to the network structure of the model. When observing an image, the human eye first looks at the global context, and then focuses attention on specific details, highlighting valuable parts and disregarding less important parts. The attention mechanism simulates this human attention by emphasizing features in regions of interest through methods such as increasing weights. SEblock [14] explicitly models the interdependencies between channels to recalibrate the feature responses of the channels. It selectively enhances useful channel features and suppresses irrelevant ones. ViT [15] divides the image into patches and inputs them into a transformer encoder with self-attention mechanism. The output is then used for classification in MLP layers, improving the ability to capture global information. CBAM [16] consists of two parts: a channel attention module and a spatial attention module. The input feature map is first passed through the channel attention module, which uses average pooling and max pooling to aggregate spatial information and generate two spatial content descriptors. These descriptors are then used to generate a channel attention map through a shared network. The CBAM mixed-domain attention mechanism module, due to its plug-and-play nature and low parameter count, has been widely applied in model improvements.

In summary, the contributions of this paper are as follows:

- We investigate a metric-based approach for industrial defect image classification in the few-shot learning scenario. The approach adopts the Classifier-Baseline framework, utilizing a deep convolutional neural network pre-trained on a large-scale dataset for image feature extraction, and calculates the cosine similarity between the images in the query set and the support set to achieve class discrimination. The classification task can be accomplished with extremely few image samples, without the need for further training adjustment for new categories.
- Innovatively, we employ DenseNet with an attention mechanism as the feature extractor. Compared to commonly used

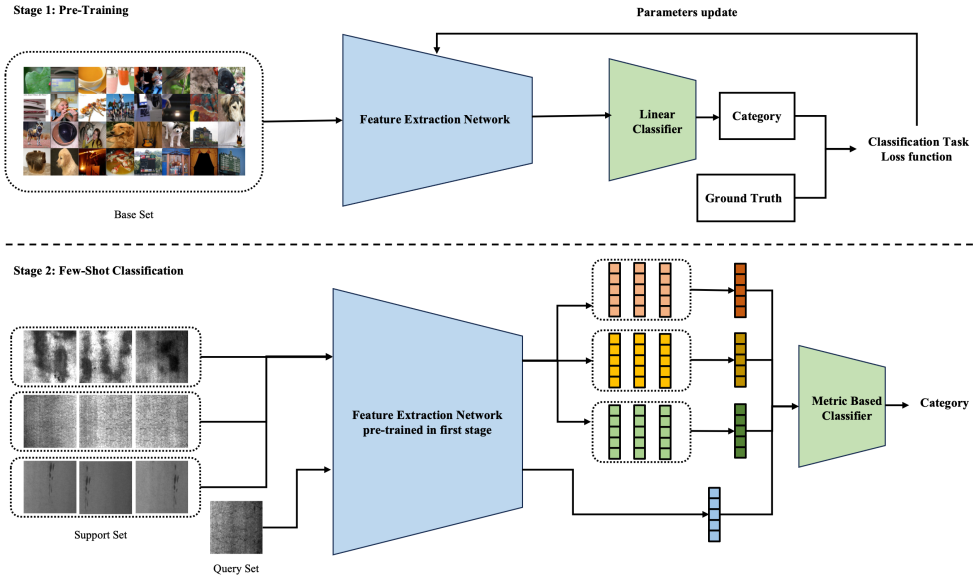


**Figure 1: Method Framework**

deep convolutional networks such as ResNet12, DenseNet allows for a lighter model overall. By incorporating a hybrid domain attention mechanism, we improve the performance and mitigate potential drawbacks caused by reducing the number of parameters.

- The effectiveness of our model is verified by applying it to actual industrial defect datasets.

Our results indicate that the method employed in this study requires very few defective samples and does not require adjustments to the model when new defect categories are introduced, making it superior to the current mainstream image classification networks when applied to real-world scenarios. Furthermore, the improved dense connected feature extraction network used in this study achieves comparable or even better detection results with only one-fourth of the parameter count of ResNet12. Additional ablation experiments confirm that the DenseNet with attention mechanism outperforms the DenseNet without attention mechanism.

In the following sections, we will first introduce the overall framework of the method in Section 2, and then provide a detailed description of the models used and the enhancements made. Subsequently, in Section 3, we will present our dataset and experimental setup, and analyze the experimental results in Section 4. Finally, in Section 5, we will summarize the article and propose potential future research directions.

## 2 METHOD

### 2.1 Framework

The framework structure of this proposed method is shown in the Figure 1, which includes two stages: Pre-Training Stage and Few-Shot Classification Stage.

In the Pre-Training Stage, a DenseNet with attention mechanism followed by a Linear classifier is used as the image classification model. A dataset with a large number of samples (such as mini-ImageNet) is used as a base category dataset at this stage, in order to train an image classification task on the classification model. The pre-trained feature extraction network can be seen as learning and acquiring the ability to extract the most representative features of images.

In the Few-Shot classification stage, the DenseNet with attention mechanism pre-trained in first stage is used as a feature extractor to extract image features. For a K-way N-shot problem, N samples from each of K categories are used to construct a support set. The images in the support set are extracted through the feature extraction network to obtain the feature vectors of each image. Then, the class center feature vector representing that category can be obtained by averaging the feature vectors of images belonging to that category. For an unknown category of query images, the same feature extraction network is used to extract image features and obtain feature vectors. The feature vectors of the query image are sequentially measured with the class center feature vectors of various categories to obtain the closest category, which is used to estimate the category to which the query image belongs.

It can be seen that in the second stage, method is metric based and does not involve model training and adjustment. Therefore, when new categories are added, inference calculations can be directly performed.

### 2.2 Feature extraction network structure

The proposed method includes three main parts: Feature Extraction Network, Linear Classifier and Metric Distance Calculation. The calculation of Linear classifier and measurement distance is

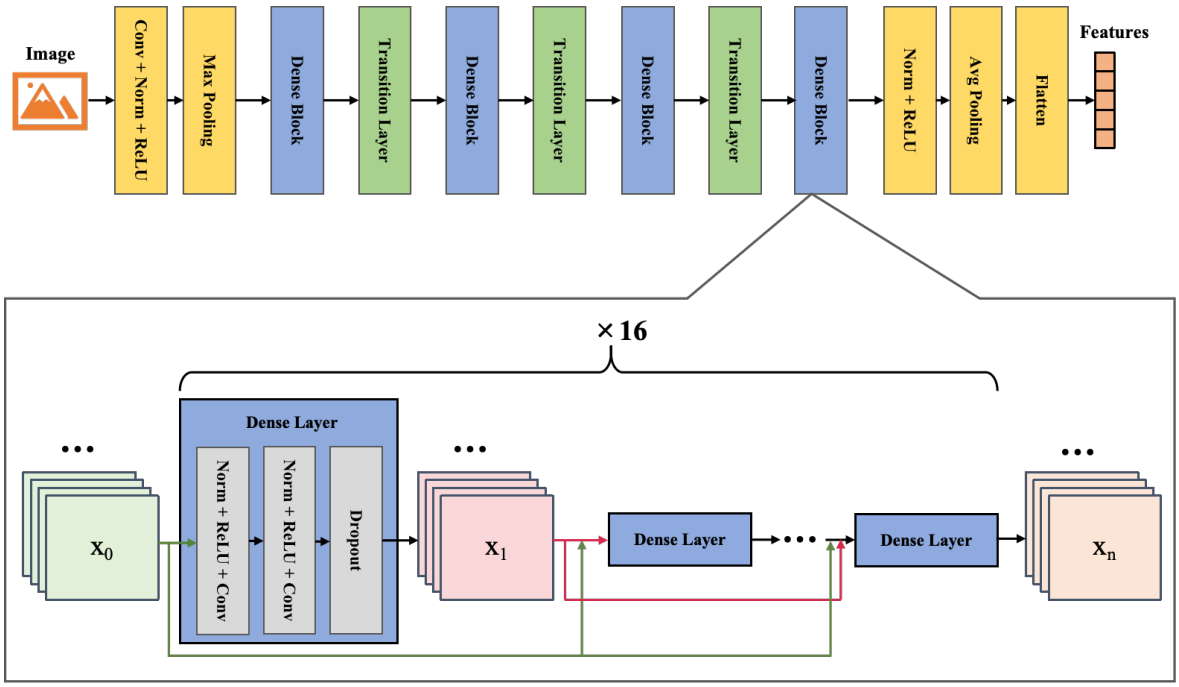


**Figure 2: Densely connected network structure**

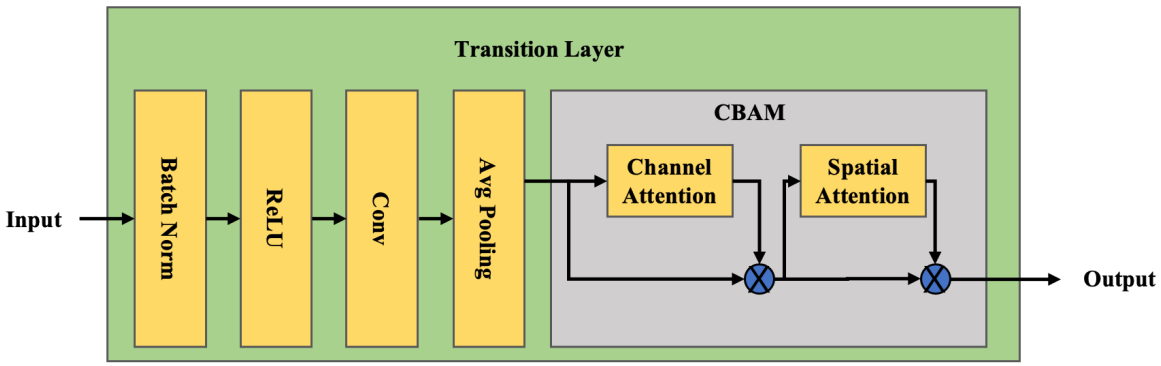


**Figure 3: Transition Layer with Attention Mechanism**

relatively simple and will not be repeated. This section mainly describes the relevant content of feature extraction network structure in detail.

*2.2.1 Densely connected network structure.* The feature extraction network adopts DenseNet with 4 Dense Blocks, each containing 16 Dense Layers. The overall structure of the feature extraction network, as well as the internal structures of Dense Block and Dense Layer, are shown in Figure 2.

In Dense Block, the connection method of Dense Layer needs to be specified. For the $l^{th}$ Dense Layer, its input includes all feature maps of the previous $l-1$ layers. Let the function of the Dense Layer be represented as $H(\cdot)$, and the $l^{th}$ Dense Layer can be represented by the following formula.

$$H\left(x_0, x_3, x_3, \cdots, x_{l-1}\right)$$

*2.2.2 Transition Layer with Attention Mechanism.* The network structure of DenseNet can extract deep features of images, but important features are not specially highlighted. Therefore, considering the different importance of each feature channel and image position, as well as the successful practice of attention mechanism in previous model improvements, we have added a mixed domain attention mechanism CBAM, which means Convolutional Block Attention Module [16], to the proposed feature extraction network. In detail, the CBAM module is added to the transition layer of DenseNet, as shown in Figure 3.

The feature map is input into the CBAM module, and the channel attention feature map is first obtained, along with important discriminative information of the channel. Then, the final attention feature map is obtained through the spatial attention mechanism. The feature map strengthened by the attention mechanism is input into the next Dense Block. The formula of CBAM is shown below.

$$CBAM(x) = x \cdot M_S(x \cdot M_C(x))$$

$$Where, \quad M_C(F) = \sigma(W_1(W_0(AvgPool(F))) + W_1(W_0(MaxPool(F))))$$

$$M_S(F) = \sigma\left(f^{7\times7}([AvgPool(F); MaxPool(F)])\right)$$

# 3 EXPERIMENT

## 3.1 Datasets

There are three Datasets used in this paper to evaluate the performance of the proposed method. mini-ImageNet, due to its wide range of categories involved, can be used to verify the universality of the proposed method. At the same time, a large portion of the categories in this dataset are also used as base category datasets to train feature extractors. The other two datasets used are defect data from actual industrial scenarios. The surface defect dataset is a publicly available dataset, collecting 6 typical defects on the surface of hot-rolled strip steel. The experiment on this dataset can demonstrate the effectiveness of our method in practical scenarios. The catenary dropper data set of high-speed rail-way is a small data set we made, which is based on the actual data of a rail-way power supply station in northwest China. Completing the classification task of defects on this dataset is also one of the motivations for the method research in this paper. However, due to the confidentiality of the data, it is not possible to disclose the data.

*3.1.1 mini-ImageNet [17].* The mini-ImageNet dataset is excerpted from the ImageNet dataset. It is a subset of the ImageNet1K dataset [18] used in the ImageNet Large Scale Visual Recognition Challenge (ILSVRC). It randomly selects 100 categories from the ImageNet1K dataset, including categories from various domains such as dogs and tires. The training set consists of 600 annotated images for each category, totaling 60,000 images. The validation set consists of 100 annotated images for each category, totaling 10,000 images. Each sample has a size of 84*84 pixels. The entire dataset is approximately 6.4GB in size, making it a suitable alternative to the complete ImageNet dataset for quick model validation, performance evaluation, and training with smaller datasets.

*3.1.2 NEU surface defect database [19].* The dataset was created by the team led by Ke-chen Song from Northeastern University. It is a dataset of surface defects on steel materials, consisting of a total of 1800 images. There are six types of defects in the dataset, namely "Crazing", "Inclusion", "Patches", "Pitted Surface", "Rolled-in Scale", "Scratches". The surface defect database released by Northeastern University (NEU) collects six typical surface defects of hot-rolled steel strips, namely rolling scale (RS), patches (Pa), cracks (Cr), pitted surface (PS), inclusions (In), and scratches (Sc). The database includes a total of 1,800 grayscale images: six different types of typical surface defects, with each defect class containing 300 samples. The dataset contains annotated data, such as class labels and defect position labels, which can be used for classification and object detection tasks. In this paper, only the class labels were used for defect image classification.

*3.1.3 dropper defect dataset.* The non-contact inspect system of the high-speed railway catenary system has been widely used in China, which can obtain regularly captured images of components. However, in some areas with low level of informatization, railway departments do not have the ability to further analyze and properly manage historical defective images. As a result, after the manual defect screening for the current quarter and subsequent maintenance work, the defect images are either deleted or stored at very low pixel resolutions in the report files. At the same time, due to the confidentiality of the data, there is no extensive data sharing among different railway bureaus. Although mainstream machine vision methods have shown excellent results, they heavily rely on large-scale data support. In this case, it becomes very difficult for railway departments in underdeveloped areas in terms of informatization to utilize advanced artificial intelligence methods for automatic defect recognition. The Dropper Defect dataset used here is sourced from the power supply section of a railway department in the northwest region of China, which includes 154 images and 41 severely deformed dropper images. Images from both categories are shown in Figure 5.

## 3.2 Experiment Settings

This paper utilized the open-source framework PyTorch for experimentation on a Linux operating system with one GPU, specifically a GTX 3080 Ti, and 12GB of memory. The mini-ImageNet dataset was divided into three parts: 64 categories for training, 16 categories for validation, and 20 categories for testing.

In the first phase, only the training and validation sets were used to train each feature extraction network. The initial learning rate was set to 0.1, with a total of 150 epochs. The learning rate was reduced by a factor of 10 at the 90th and 110th epochs. The SGD algorithm was employed as the optimizer, with a weight decay coefficient of 0.0005. The update parameter criterion utilized the cross-entropy loss between the predicted and true labels. Convergence was achieved by the end of the first phase.

In the second phase, three different datasets were used for experimentation: the mini-ImageNet testing set with 20 categories, the NEU surface defect database, and the dropper defect dataset. Cosine similarity was employed as the metric. For each feature extractor and dataset, query set contains 15 images, experiments were conducted using 5-way 1-shot, 5-way 5-shot, 2-way 1-shot, and 2-way 5-shot. The dropper defect dataset only contained two categories, hence only 2-way 1-shot and 2-way 5-shot experiments were conducted.

# 4 RESULTS AND DISCUSSION

## 4.1 Comparative Experiments

ResNet-12 is utilized as a baseline for comparing the model's performance. The significance of parameter count is emphasized in the analysis of experimental results. In this study, the improved DenseNet has a parameter count of 2M, while ResNet12 has a parameter count of 8M. The proposed feature extraction network in this paper has only one-fourth of the parameter count compared

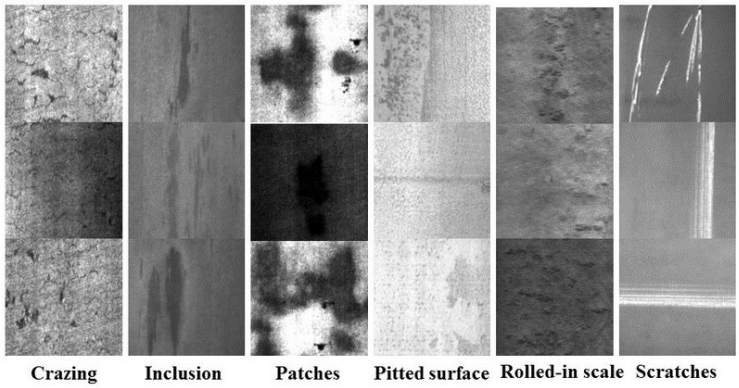


**Figure 4: Examples from NEU surface defect database[19]**

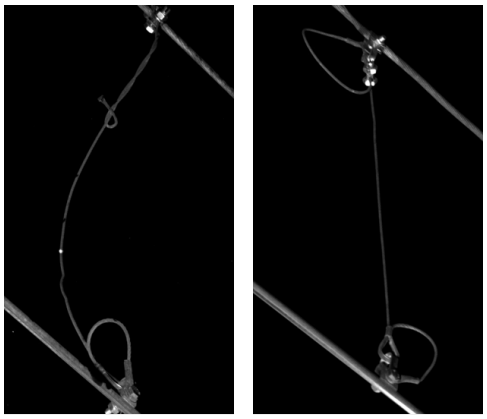


**Figure 5: Examples from Dropper Defect Dataset**

to the baseline. Therefore, in analyzing the results, it is essential to examine whether the model can still achieve performance comparable to or even surpassing the baseline's performance when significantly reducing the parameter count.

Experiments were conducted on three different datasets: mini-ImageNet, surface defect, and dropper defect. These experiments aimed to validate the generalizability of the proposed approach and its performance on real industrial datasets. The experimental results are presented in the Table 1.

It can be observed that our method achieves accuracy levels close to models with higher parameter quantities when conducting 5-way 5-shot, 5-way 1-shot, and 2-way 1-shot tasks on mini-ImageNet. Additionally, our method surpasses the reference baselines in the relatively simpler task of 5-way 1-shot.

The experimental results on the surface defect dataset collected in real industrial scenarios indicate that our method approaches the accuracy of the baseline model in the 1-shot task, and even outperforms the baseline model by one percentage point in tasks like 5-shot that have a greater number of reference samples. The experimental results on the dropper defect dataset also demonstrate the effectiveness of our method in different practical scenarios.

## 4.2 Ablative Experiments

In order to investigate the impact of incorporating attention mechanisms on the performance of the feature extraction network, we conducted ablative experiments to compare the differences in results before and after incorporating attention mechanism. The baseline model for feature extraction in all experiments was DenseNet, with the only variation being the inclusion or exclusion of CBAM for improvement. The dataset and parameter settings were consistent throughout the experiments. The experimental results are presented in Table 2.

The results of the ablative experiment indicate that the addition of an attention mechanism to the feature extractor, compared to using DenseNet as the feature extractor without an attention mechanism, leads to higher accuracy in all tasks. The improved model can achieve an average increase in accuracy of 1 to 2 percentage points,

**Table 1: Classification Accuracy of Comparative Experiments (%)**

| Dataset | Feature Extractor | 5-way 5-shot | 5-way 1-shot | 2-way 5-shot | 2-way 1-shot |
|---|---|---|---|---|---|
| mini-ImageNet | Resnet12 | 78.21 | 59.49 | 87.40 | 81.35 |
| | DenseNet-CBAM | 74.39 | 56.24 | 89.26 | 80.05 |
| Surface defect | Resnet12 | 88.41 | 75.75 | 94.63 | 90.99 |
| | DenseNet-CBAM | 89.78 | 74.10 | 95.12 | 90.70 |
| Dropper defect | Resnet12 | | | 79.85 | 82.67 |
| | DenseNet-CBAM | | | 77.48 | 78.29 |

**Table 2: Classification Accuracy of Ablative Experiments (%)**

| Dataset | CBAM | 5-way 5-shot | 5-way 1-shot | 2-way 5-shot | 2-way 1-shot |
|---|---|---|---|---|---|
| mini-ImageNet | √ | 74.39 | 56.24 | 89.26 | 80.05 |
| | | 73.41 | 55.15 | 87.30 | 78.82 |
| Surface defect | √ | 89.78 | 74.10 | 95.12 | 90.70 |
| | | 87.79 | 71.98 | 93.02 | 88.39 |
| Dropper defect | √ | | | 77.48 | 78.29 |
| | | | | 71.73 | 73.24 |

and even a significant improvement of almost 5.7 percentage points in the 2-way 5-shot task on the dropper defect dataset. This shows that incorporating attention mechanisms can enhance the feature extraction ability of DenseNet in Few-Shot image classification tasks, and demonstrates the effectiveness of the improvements proposed in this paper.

## 5 CONCLUSION

In conclusion, this paper makes several contributions to the field of industrial defect image classification in the few-shot learning scenario. Firstly, we propose a metric-based approach using the Classifier-Baseline framework, which leverages a deep convolutional neural network pre-trained on a large-scale dataset for image feature extraction. By calculating the cosine similarity between query and support images, our approach achieves effective class discrimination with minimal training samples and without the need for additional adjustment for new defect categories.

Secondly, we innovate by adopting DenseNet with an attention mechanism as the feature extractor. Compared to popular networks like ResNet12, DenseNet offers a lighter overall model. Through the incorporation of a hybrid domain attention mechanism, we improve performance and mitigate potential drawbacks caused by reducing the number of parameters.

Thirdly, our proposed model is extensively evaluated on actual industrial defect datasets, confirming its effectiveness in real-world scenarios.

The results demonstrate that our method requires minimal defective samples and eliminates the need for model adjustments when introducing new defect categories, thus outperforming mainstream image classification networks in practical applications. Additionally, the improved dense connected feature extraction network used in our study achieves comparable or even better detection results while using only one-fourth of the parameter count of ResNet12. Furthermore, ablative experiments confirm that DenseNet with an attention mechanism outperforms DenseNet without an attention mechanism.

Overall, our work contributes to the advancement of few-shot learning in industrial defect image classification, providing a robust and efficient solution for real-world scenarios.

Future research directions can further enhance the field of industrial defect image classification in the few-shot learning scenario. One potential direction is to explore the effectiveness of transfer learning techniques in this context. By leveraging pre-trained models from related domains or datasets, researchers can potentially improve the classification performance even further.

Furthermore, incorporating more advanced attention mechanisms, such as self-attention or transformer-based models, could be explored to enhance the feature extraction capabilities of the network. These attention mechanisms have demonstrated impressive performance in various computer vision tasks and may provide additional discriminative power for industrial defect image classification.

Additionally, exploring the combination of different modalities, such as incorporating textual or sensor data in conjunction with visual information, could lead to more comprehensive and robust defect classification systems. This multi-modal approach may capture diverse aspects of industrial defects, leading to improved accuracy and reliability.

In summary, future research in the field of industrial defect image classification in the few-shot learning scenario can focus on the fusion of cross-domain and multi-modal information and enhancement of model feature extraction capabilities aiming to advance the state-of-the-art further and address the challenges in real-world scenarios.

## ACKNOWLEDGMENTS

We would like to express our gratitude to the relevant departments that have provided us with dropper defect data samples of high-speed railway catenary system.

## REFERENCES

[1] LeCun, Y., Bottou, L., Bengio, Y., & Haffner, P. (1998). Gradient-based learning applied to document recognition. Proceedings of the IEEE, 86(11), 2278-2324.
[2] Krizhevsky, A., Sutskever, I., & Hinton, G. E. (2012). Imagenet classification with deep convolutional neural networks. Advances in neural information processing systems, 25.
[3] Simonyan, K., & Zisserman, A. (2014). Very deep convolutional networks for large-scale image recognition. arXiv preprint arXiv:1409.1556.
[4] Szegedy, C., Liu, W., & Jia, Y. Going deeper with convolutions 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR) June 2015Boston. MA, USA1–9, 10.
[5] He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep residual learning for image recognition. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 770-778).
[6] Wang, Y., Yao, Q., Kwok, J. T., & Ni, L. M. (2020). Generalizing from a few examples: A survey on few-shot learning. ACM computing surveys (csur), 53(3), 1-34.
[7] Chen, Y., Liu, Z., Xu, H., Darrell, T., & Wang, X. (2021). Meta-baseline: Exploring simple meta-learning for few-shot learning. In Proceedings of the IEEE/CVF international conference on computer vision (pp. 9062-9071).
[8] Iandola, F. N., Han, S., Moskewicz, M. W., Ashraf, K., Dally, W. J., & Keutzer, K. (2016). SqueezeNet: AlexNet-level accuracy with 50x fewer parameters and< 0.5 MB model size. arXiv preprint arXiv:1602.07360.
[9] Chollet, F. (2017). Xception: Deep learning with depthwise separable convolutions. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 1251-1258).
[10] Howard, A. G., Zhu, M., Chen, B., Kalenichenko, D., Wang, W., Weyand, T., ... & Adam, H. (2017). Mobilenets: Efficient convolutional neural networks for mobile

vision applications. arXiv preprint arXiv:1704.04861.
[11] Türkmen, S., & Heikkilä, J. (2019, May). An efficient solution for semantic segmentation: Shufflenet v2 with atrous separable convolutions. In Scandinavian Conference on Image Analysis (pp. 41-53). Cham: Springer International Publishing.
[12] Huang, G., Liu, Z., Van Der Maaten, L., & Weinberger, K. Q. (2017). Densely connected convolutional networks. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 4700-4708).
[13] Pleiss, G., Chen, D., Huang, G., Li, T., Van Der Maaten, L., & Weinberger, K. Q. (2017). Memory-efficient implementation of densenets. arXiv preprint arXiv: 1707.06990.
[14] Hu, J., Shen, L., & Sun, G. (2018). Squeeze-and-excitation networks. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 7132-7141).
[15] Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., ... & Houlsby, N. (2020). An image is worth 16x16 words: Transformers for image recognition at scale. arXiv preprint arXiv:2010.11929.
[16] Woo, S., Park, J., Lee, J. Y., & Kweon, I. S. (2018). Cbam: Convolutional block attention module. In Proceedings of the European conference on computer vision (ECCV) (pp. 3-19).
[17] Vinyals, O., Blundell, C., Lillicrap, T., & Wierstra, D. (2016). Matching networks for one shot learning. Advances in neural information processing systems, 29.
[18] Deng, J., Dong, W., Socher, R., Li, L. J., Li, K., & Fei-Fei, L. (2009, June). Imagenet: A large-scale hierarchical image database. In 2009 IEEE conference on computer vision and pattern recognition (pp. 248-255). Ieee.
[19] He, Y., Song, K., Meng, Q., & Yan, Y. (2019). An end-to-end steel surface defect detection approach via fusing multiple hierarchical features. IEEE transactions on instrumentation and measurement, 69(4), 1493-1504.