Proceedings of CONF-MLA 2025 Symposium: Applied Artificial Intelligence Research
DOI: 10.54254/2755-2721/2025.BJ25449

Evaluating the Detection Performance of Different Convolutional Neural Network Models for Steel Surface Defects

Zhichao Chen

School of Optical Science and Engineering, Zhejiang University, Hangzhou, China
3200103703@zju.edu.cn

Abstract: To ensure quality control in the steel industry, accurate and efficient surface defect detection technology is essential. Deep learning, especially Convolutional Neural Networks (CNNs), has shown great potential in this field, but there is still a lack of systematic performance comparison of models with different architectural concepts under a unified framework. This study comprehensively evaluates four representative CNN models - VGG16, ResNet50, InceptionV3, and MobileNetV2 - based on the steel surface defect dataset publicly available from Northeastern University (NEU-CLS). This study adopts a unified data preprocessing, data augmentation strategy, and a fine-tunable unified transfer learning method when conducting experiments. The performance of the models is then evaluated by comparing the accuracy, precision, recall, F1 score, and confusion matrix of each model. The experimental results show that all models have very good performance in the task of steel surface defect classification, with test accuracy exceeding 97%. Among all models, VGG16 has the highest accuracy of 99.26%. The lightweight MobileNetV2 has the same accuracy as the more complex ResNet50, both reaching 98.52%. This result shows that MobileNetV2 has achieved a good balance between performance and computational efficiency. In short, the main contribution of this study is to provide a judgment benchmark for future model selection. When weighing the model complexity and classification accuracy, you can make a choice based on the results of this experiment.

Keywords: Defect detection, deep learning, convolutional neural networks, model comparison

1. Introduction

Steel plates are the basic materials of modern industry, and their surface quality is crucial to product performance, reliability, and downstream processing. With the continuous development of industrial automation, the demand for high-precision surface defect detection has become increasingly prominent, and efficient, accurate, strong practical and theoretical detection technologies are urgently needed [1]. However, traditional detection methods (such as manual visual inspection) are highly subjective, inefficient, and difficult to scale. Physical methods such as eddy current and

DOI: 10.54254/2755-2721/2025.BJ25449

infrared detection are also limited by cost, efficiency, and applicability to different defect types and working conditions [1].

However, steel surface defects present a variety of patterns, such as cracks, pits, inclusions, and scratches, with fuzzy boundaries and significant intra-class differences. In addition, challenges such as illumination variations and reflection interference in industrial environments also complicate accurate detection [2]. Although the technology based on traditional machine learning (such as K-nearest neighbor algorithm) has taken a step towards automation [3], it relies on manually designed feature extractors and its accuracy is difficult to meet the requirements when faced with complex and changing industrial scenarios [4]. In contrast, deep learning, especially Convolutional Neural Networks (CNNs), provides end-to-end learning and automatic feature extraction, which significantly improves the defect detection capabilities of materials such as steel plates and wood boards [5, 6]. CNNs do not require manually constructed rules and are able to learn deep hierarchical representations from raw images.

In image classification tasks, traditional deep networks like VGGNet and ResNet have demonstrated exceptional performance [7, 8]. ShuffleNet builds effective networks using grouped convolution and channel reordering, while certain lightweight neural network design approaches, including MobileNetV1 and MobileNetV2 [9, 10], employ depthwise separable convolution [11]. While retaining good accuracy, these lightweight models significantly lower the computational complexity and model size.

Although existing studies have verified the effectiveness of certain heavyweight or lightweight models in steel defect classification tasks, there is still a lack of a systematic and comprehensive performance comparison of these representative CNN architectures with different design philosophies (high accuracy vs. high efficiency) under a unified experimental framework. The different datasets, preprocessing processes, and evaluation metrics used in different studies make it difficult to make a fair horizontal comparison of the performance of various models, and also blur the inherent trade-off between model complexity and recognition performance.

Therefore, this study aims to conduct an empirical evaluation of representative CNN models (including InceptionV3, MobileNetV2, ResNet50, and VGG 16) on the publicly available NEU-CLS dataset. It unifies the preprocessing, data augmentation, and training settings, adopts multiple evaluation metrics (accuracy, precision, recall, F1 score, and confusion matrix, etc.), and explores the trade-off between performance and complexity. Its goal is to provide theoretical insights and practical guidance for model selection and deployment in industrial defect detection systems.

2. Method

2.1. Dataset preparation

This study makes use of Northeastern University's publicly available steel surface defect dataset, known as NEU-CLS [12]. This dataset was created specifically to serve as a benchmark for the classification task of common surface flaws in hot-rolled steel strips.

There are 1,800 grayscale photos with a 200×200 pixel resolution in the NEU-CLS dataset. With precisely 300 samples in each of the six typical fault types, the dataset has been meticulously crafted to attain class balance. Crazing, Inclusion, Patches, Pitted Surface, Rolled-in Scale, and Scratches are the six categories.

Fig. 1 shows representative image samples of the six defect categories in the dataset, intuitively presenting the visual characteristics of various defects.

DOI: 10.54254/2755-2721/2025.BJ25449

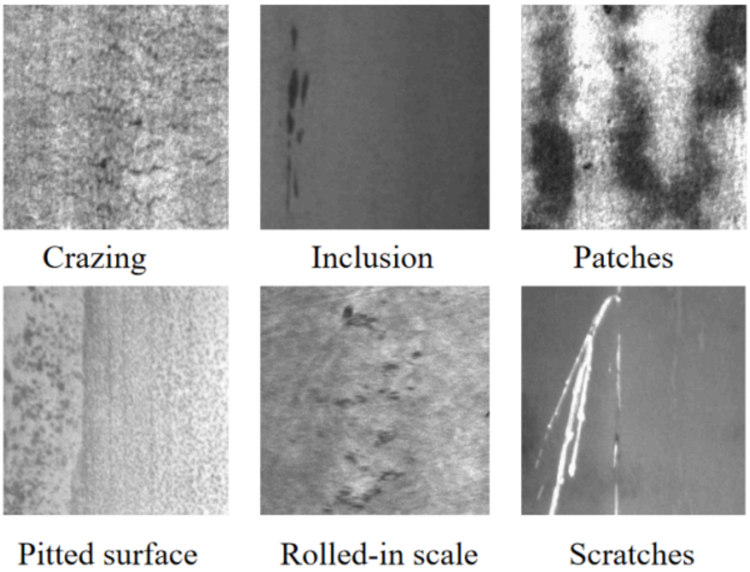

Figure 1. Samples in the NEU-CLS dataset [12]

This study created a systematic data processing and enhancement procedure to guarantee the validity of the experimental findings and the rigor of the research methodology.

This study rigorously separated the dataset into three mutually exclusive subsets: the training set, validation set, and test set, using a division ratio of 70%/15%/15% in order to obtain an objective assessment of the model's capacity for generalization. This approach guarantees that the model's final performance is assessed using entirely unseen data. The results of the particular division are as follows:

All images are processed using the ImageDataGenerator tool of the TensorFlow Keras framework. The preprocessing process aims to make the data meet the input requirements of the pre-trained model.

To match the input specifications of the CNN model pre-trained on ImageNet, all 200×200 pixel grayscale images are uniformly resized to 224×224 pixels when loaded and converted from single-channel grayscale images to three-channel RGB images.

The pixel values of the image are scaled from the original [0, 255] integer range to the [0, 1] floating point range.

In order to expand the effective training sample size, reduce possible overfitting of the model, and improve the robustness and generalization ability of the model, only a series of online data augmentation techniques were implemented on the training set data. These techniques are designed to simulate possible changes in images in reality while avoiding the introduction of artifacts that may mask key defect features.

Procedure for random rotation: 1) Rotate the image at random within a ±20 degree range. 2) Alternate the image at random in both horizontal and vertical directions, moving no more than 10% of its width or height. 3) With a 50% chance, flip the picture horizontally. 4) The closest neighbor interpolation technique is applied to fill in the blank spaces left by the transformation.

During the experiment, no data augmentation operations were performed on the validation set and the test set, only normalization was performed to ensure the objectivity and consistency of the model evaluation.

DOI: 10.54254/2755-2721/2025.BJ25449

2.2. Cnn-based classification models

This study selected four CNN architectures that are milestones in the field of computer vision and applied them to the steel surface defect classification task using the transfer learning paradigm. These models were pre-trained on the ImageNet dataset, and their unique architecture designs are described as follows.

2.2.1. Inception V3

The core idea of this model is to efficiently expand the network through decomposition convolution and design principles, while effectively reducing computational complexity without sacrificing receptive field [13, 14]. Its Inception module uses convolution kernels of various sizes at the same time and introduces asymmetric convolution (for example, decomposing n x n into 1 x n and n x 1), so that the network can efficiently extract features of multiple scales and achieve a good balance between computational efficiency and model accuracy.

2.2.2. MobileNetV2

This model is intended for embedded and mobile platforms with constrained processing power [11]. Its main innovation is the use of linear bottlenecks and inverted residuals to achieve great efficiency on mobile terminals. It is incredibly lightweight, quick, memory-efficient, and accurate on mobile terminals thanks to the modular design of "expansion-depthwise separable convolution-compression" and the use of linear activation functions in the bottleneck layer to preserve feature information.

2.2.3. ResNet50

The introduction of the residual network (ResNet) solves the "degradation" problem in the training process of deep neural networks [8]. Its core is the deep residual learning framework. By introducing the "shortcut connection", the network learns the residual function F(x) instead of the direct target mapping H(x). The "bottleneck" residual block in ResNet50 adopts the "1x1, 3x3, 1x1" convolution structure, which not only ensures the depth but also controls the computational cost, making it possible to train a network with hundreds of layers.

2.2.4. VGG16

The design philosophy of the VGG model is simplicity and depth [7]. It proves that building a deep network by simply and uniformly stacking small-size (3x3) convolution kernels is an effective way to improve performance. By joining several 3x3 convolution layers in succession, VGG16's regular structure eliminates the need for large-size convolution kernels while simultaneously increasing the network's capacity for nonlinear expression and lowering its parameter count. Because of its strong feature extraction capabilities, it is frequently utilized as a foundational model for further study.

2.3. Implementation details

To ensure the objectivity of the comparison, all models were trained using a unified transfer learning strategy under the TensorFlow Keras framework. Each model loaded the ImageNet pre-trained

DOI: 10.54254/2755-2721/2025.BJ25449

weights, first freezing the backbone network and training the newly added classification head, then unfreezing some convolutional layers and fine-tuning with an extremely low learning rate (1e-5).

Every model employed a batch size of 16, the Adam optimizer, and the Categorical Crossentropy loss function. To avoid overfitting, EarlyStopping was applied throughout the training phase. Accuracy, precision, recall, F1 score, and confusion matrix were all used to thoroughly assess the final model's performance on a separate test set.

3. Results and discussion

This section evaluates the performance of four CNN models on the NEU-CLS steel surface defect dataset. This evaluation dataset comes from an independent test set that was previously separated.

3.1. Overall performance comparison

As indicated in Table 1, the models are evaluated using the performance metrics of accuracy, precision, recall, and F1 score. Table 1 shows that all of the networks perform quite well overall, with a test accuracy of above 97%. This outcome demonstrates how well transfer learning from established CNN architectures works for jobs involving industrial visual inspection.

Table 1. Model performance comparison on NEU-CLS dataset

	Accuracy	Precision	Recall	F1-Score
MobileNetV2	0.9852	0.9855	0.9852	0.9852
VGG16	0.9926	0.9929	0.9926	0.9926
ResNet50	0.9852	0.9853	0.9852	0.9851
InceptionV3	0.9778	0.9780	0.9778	0.9777

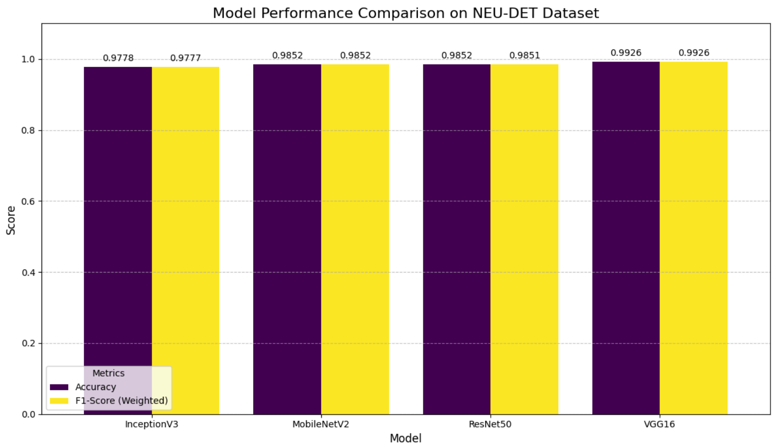

Figure 2. Comparison of overall accuracy and weighted F1-score (picture credit : original)

Among the four models shown in Fig. 2, VGG16 performs well in all evaluation indicators, with an overall accuracy of 99.26% and a weighted F1 score of 0.9926. This shows that its sequential architecture consisting of multiple 3×3 convolutional layers connected in series can accurately capture the characteristics of the six defect categories in NEU-CLS, and the effect is very good.

Both MobileNetV2 and ResNet50 achieved an overall accuracy of 98.52%, with only slight differences in precision and recall. However, there are huge differences in the architectural

DOI: 10.54254/2755-2721/2025.BJ25449

complexity and computational requirements. This result shows that lightweight architectures also have the potential to achieve high-performance classification in some resource-constrained situations.

Meanwhile, although InceptionV3 also showed good performance, it had the lowest accuracy of all models, only 97.78%. This relatively low data shows that the feature selection method of InceptionV3 is not the best choice for this dataset.

3.2. Performance and error analysis of each category

In addition, the confusion matrix (Fig. 3, Fig. 4, Fig. 5 and Fig. 6) was used to check the detection performance of the models for each category. As can be seen from the figure, all models performed well, with excellent classification capabilities for both patch and rolling scale categories, and usually achieved perfect precision and recall.

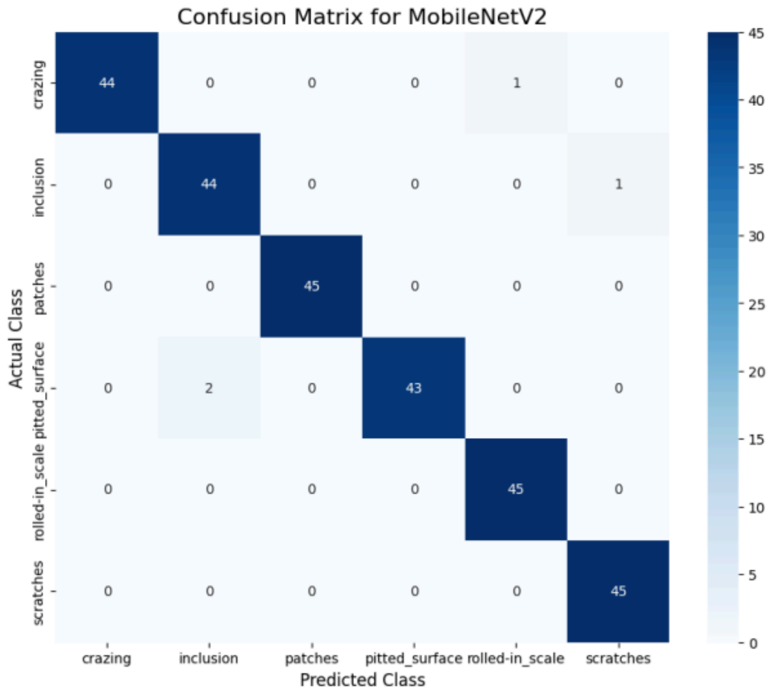

Figure 3. Confusion matrices for MobileNetV2 (picture credit : original)

DOI: 10.54254/2755-2721/2025.BJ25449

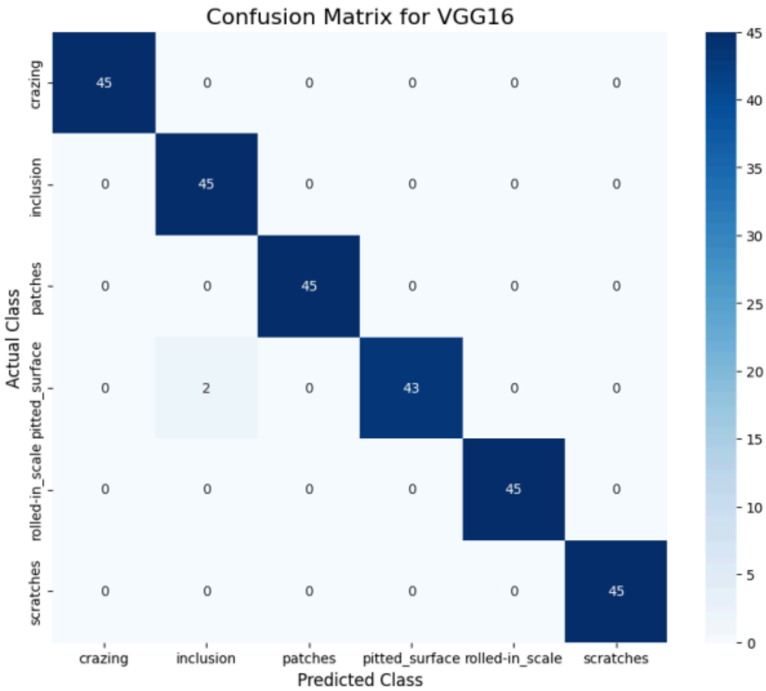

Figure 4. Confusion matrices for VGG16 (picture credit : original)

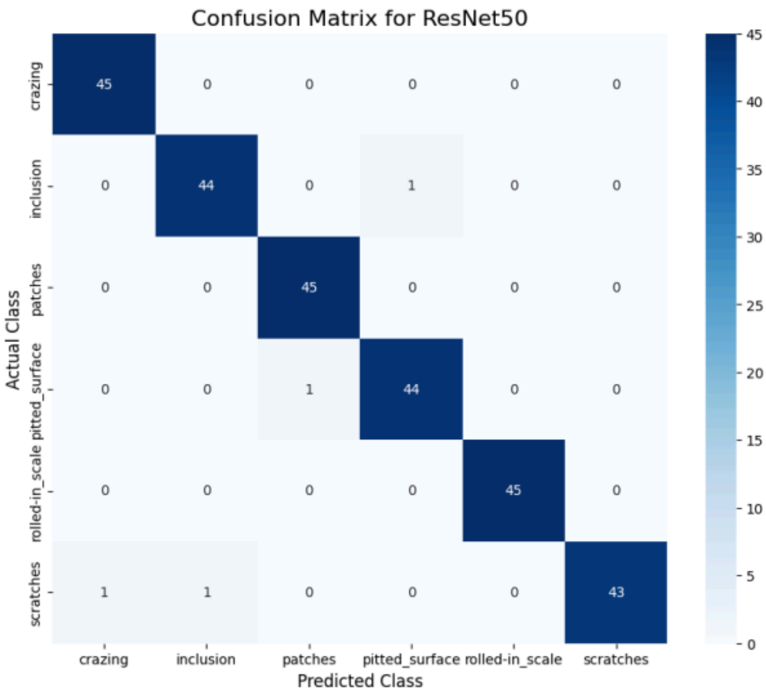

Figure 5. Confusion matrices for Resnet50 (picture credit : original)

DOI: 10.54254/2755-2721/2025.BJ25449

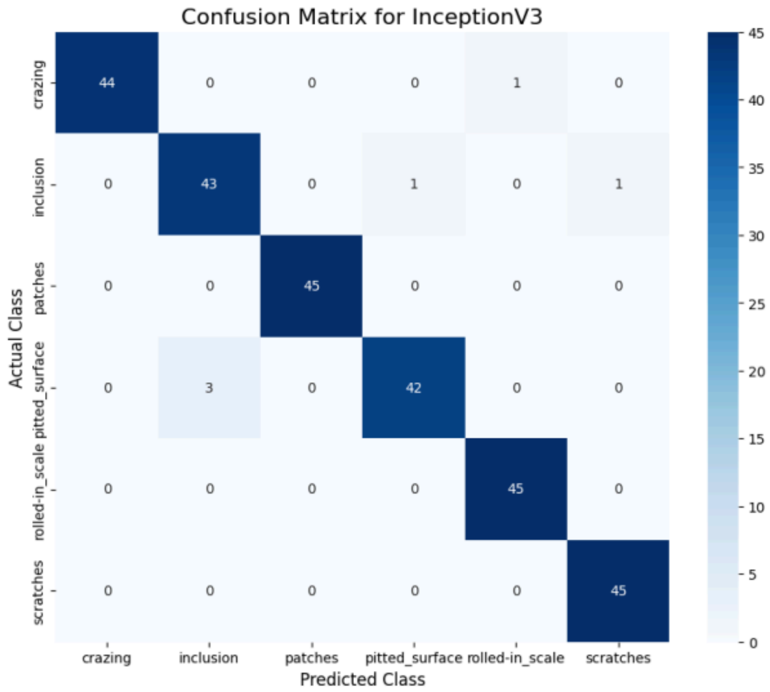

Figure 6. Confusion matrices for InceptionV3 (picture credit : original)

The most misclassification among all models was the confusion between the pitted surface (pitted_surface) and inclusion (interclusion) categories. Visually, the two usually appear as small local defects with overlapping texture features, which have certain similarities. VGG16 misclassified two pitted_surface images as inclusions. InceptionV3 misclassified three pitted_surface samples and one inclusion sample. MobileNetV2 also misclassified this category twice.

These findings highlight a potential area for improvement, perhaps by training and strengthening the model's classification ability for these categories through methods such as enhanced data augmentation, domain-specific preprocessing, or introducing attention mechanisms that prioritize fine-grained details.

Among all models, VGG16 performed best with only two errors in the entire test set. ResNet50 had a different error pattern from the other models, misclassifying scratch images as cracks and inclusions respectively. This different error suggests that there may be problems in modeling linear defect patterns, which requires the use of more targeted feature representations.

3.3. Model complexity and its practical implications

Although VGG16 has the highest classification accuracy among all the tested models, it also has the highest computational complexity, with a model parameter count of about 138 million. In contrast, MobileNetV2 has only about 3.5 million parameters, while its classification accuracy is still good. This feature makes it an ideal candidate for deploying similar tasks on resource-constrained systems.

The results of this study confirm that the lightweight architecture in MobileNetV2 does not affect classification accuracy for the NEU-CLS dataset. These results provide valuable suggestions for model selection for deep learning in industrial deployment scenarios. The most appropriate model should be selected based on specific needs, while considering the balance between accuracy and computational efficiency. When the main goal is maximum classification accuracy and computing

DOI: 10.54254/2755-2721/2025.BJ25449

resources are sufficient, VGG16 provides a reliable and high-performance solution. For power-constrained scenarios, MobileNetV2 may be a better choice.

4. Conclusion

This study systematically compares the classification capabilities of four mainstream convolutional neural network architectures - InceptionV3, MobileNetV2, ResNet50 and VGG16 - on the NEU-CLS steel surface defect dataset.

In this experiment, all models performed well, with an accuracy of more than 97%, which proves the effectiveness and applicability of deep learning methods in industrial defect detection tasks. Among all models, VGG16 performed the best, with an accuracy of 99.26%. The lightweight MobileNetV2 model has the same accuracy as the more complex ResNet50 (98.52%), but the MobileNetV2 model significantly reduces the computational complexity. The high similarity in visual features between the "pitted surface" and "inclusion" categories makes the main errors of each model concentrated between these two.

In summary, VGG16 is more suitable for application scenarios that prioritize accuracy; MobileNetV2 is more suitable in industrial environments with limited resources or real-time deployment.

Based on the results of this study, there are still some directions for future research. First, this study focuses on classic CNN architecture, and future research can use more architecture and incorporate emerging architectures, especially vision transformers (ViT) and their efficient variants. Second, although the NEU-CLS dataset provides a standardized benchmark, future research should focus on validating these findings on more realistic industrial datasets, which will ensure that the developed solutions can be generalized to actual deployment.

References

[1] Li, L. (2015) The UAV Intelligent Inspection of Transmission Lines. Proceedings of the 2015 International Conference on Advances in Mechanical Engineering and Industrial Informatics, 2015.

[2] Liu, Y., et al. (2020) A Light-Weight Deep-Learning Model with Multi-Scale Features for Steel Surface Defect Classification. Materials, 13(20), 4629.

[3] Shamsi, M. and Beheshti, S. (2025) Separability and Scatteredness (S&S) Ratio-Based Efficient SVM Regularization Parameter, Kernel, and Kernel Parameter Selection. Pattern Analysis and Applications, 28(1), 33.

[4] Zhao, L.Y. and Wu, Y.Q. (2022) Research Progress of Surface Defect Detection Methods Based on Machine Vision. Chinese Journal of Scientific Instrument, 43(1), 198-219.

[5] LeCun, Y., et al. (1998) Gradient-Based Learning Applied to Document Recognition. Proceedings of the IEEE, 86(11), 2278-2324.

[6] He, Y., et al. (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.

[7] Simonyan, K. and Zisserman, A. (2014) Very Deep Convolutional Networks for Large-Scale Image Recognition. arXiv Preprint, arXiv: 1409.1556.

[8] He, K., et al. (2016) Deep Residual Learning for Image Recognition. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016.

[9] Jeon, Y.J., et al. (2015) Steel-Surface Defect Detection Using a Switching-Lighting Scheme. Applied Optics, 55(1), 47-57.

[10] Andrew, G. and Zhu, M. (2017) Efficient Convolutional Neural Networks for Mobile Vision Applications. MobileNets, 10(2), 151.

[11] Sandler, M., et al. (2018) MobileNetV2: Inverted Residuals and Linear Bottlenecks. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018.

[12] Bao, Y., Song, K., Liu, J., Wang, Y., Yan, Y., Yu, H. and Li, X. (2021) Triplet-Graph Reasoning Network for Few-Shot Metal Generic Surface Defect Segmentation. IEEE Transactions on Instrumentation and Measurement, 70, 1-

DOI: 10.54254/2755-2721/2025.BJ25449

1.
[13] Szegedy, C., et al. (2016) Rethinking the Inception Architecture for Computer Vision. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016.
[14] Szegedy, C., et al. (2017) Inception-v4, Inception-ResNet and the Impact of Residual Connections on Learning. Proceedings of the AAAI Conference on Artificial Intelligence, 31(1).