Data Engineering and Applications

Research Paper

Deep Learning-Based Weld Seam Quality Assessment Using a Joint ResNeXt101 and AdaBound Model

Adrian Kwiatkowski[1], Natalia Gawlikowska[1,] and Konrad Mariusz Pawlak[1,*]

[1] Maria Curie-Sklodowska University, Faculty of Mathematics, Physics and Computer Science, 20-031 Lublin, Poland
*Corresponding author: knorad.mp@wpias.edu.pl

Abstract. Due to the application of computer vision technology, weld defect detection has become relatively reliable and efficient in the industrial field. By combining the ResNeXt101 backbone with the AdaBound adaptive optimizer, a hybrid deep learning model was constructed to achieve long-term effective welding quality assessment. The techniques used to enhance the effectiveness of comprehensive data augmentation are grouped convolution feature extraction and dynamic learning rate adjustment. In the experiment, two large-scale, multi-domain weld seam datasets were used, totaling 21,000 labeled samples. In the positive experiments, the proposed hybrid model outperformed the standard ResNet50 and DenseNet121 baselines, achieving an accuracy of 97.1%, a macro F1 score of 95.0%, and a macro-AUC of 99.2%. Strict ablation studies indicate that the choice of different backbone networks, optimizers, augmentation sets, and input resolutions all have a certain impact on performance improvement. The aforementioned robustness tests indicate that it has good generalization ability for cross-domain data, relatively strong noise tolerance, and balanced performance in detecting both common and rare defect types. The segmentation results indicate that over 92% of the samples have an Intersection over Union (IoU) greater than 0.8, thus they are of high quality. According to the above research results, the hybrid deep learning method has achieved good performance in automatic welding quality inspection. These findings also indicate that the application of this method in large-scale manufacturing has empirical significance.

Keywords: *Deep Learning, Weld Defect Detection, Grouped Convolution, Adaptive Optimization, Data Augmentation, Industrial Inspection, Robustness*

Received on 17 October 2025, Accepted on 23 December 2025, Published on 5 Jan2026

Introduction

The quality of welds directly affects the safety and lifespan of metal components in the automotive, aerospace, power plant, and other industries [1]. Ensure weld quality to avoid serious defects and high maintenance costs in these maintenance-prone areas [2]. In recent years, three main traditional welding inspection methods—radiographic testing, ultrasonic testing, and visual inspection—have been widely used [3]. The labor intensity of manual and semi-automatic inspections is very high, making human errors likely; moreover, they are not as reliable and repeatable in large industrial environments [4]. For noisy environments, irregular shapes, or small-scale defects, traditional machine vision and early computational detection methods often struggle to handle the complex visual variations of real weld seams [5]. Shallow machine learning techniques have shown improvements in defect recognition and classification accuracy, but poor adaptability and scalability when handling large amounts of data and complex problems remain one of the issues these technologies face [6]. Compared to traditional manually designed descriptors, Convolutional Neural Networks (CNNs) have improved the ability to extract multi-scale features. However, the early applications of CNNs were limited by insufficient depth and a lack of complex architectural improvements [7]. Therefore, welding inspection requires next-generation automation technology [8].

With the latest advancements in deep learning, many automated quality inspections have already entered the commercial sector [9]. ResNet and its derivative networks are a type of deeper convolutional neural network

(CNN). It enhances the model's expressive and learning capabilities in large-scale image recognition by adding residual connections and multiple branches [10]. The ResNeXt101 network uses grouped convolutions and cardinality to achieve an efficient, scalable, and highly modular structure. It performs excellently in industrial detection under various operational conditions [11]. On the other hand, as such networks become increasingly complex, there is now a need to improve the adaptive optimization algorithms for model training. AdaBound is a new proposal suitable for modern deep networks, dynamically switching between the generalization bounds of SGD and the fast convergence speed of Adam [12]. According to recent research, using adaptive optimizers and high-capacity architectures can improve the training speed and generalization ability of models [13]. However, integrating the aforementioned techniques into a comprehensive end-to-end system for robust weld quality assessment remains an unresolved issue. Serious issues regarding cross-domain adaptability, inference speed, and interpretability remain unresolved [14]. In the near future, automated systems will be introduced to address the aforementioned issues [15].

This paper proposes a deep learning-based weld quality assessment method, addressing the aforementioned issues by integrating ResNeXt101 and the AdaBound optimizer into an end-to-end architecture. Our contributions include the development of a hybrid system that combines advanced feature extraction and adaptive optimization, extensive validation on controlled and real industrial datasets, and detailed analysis of robustness, ablation behavior, and potential deployment scenarios.

Principles of Model Design

Feature Learning with Advanced Convolutional Architectures

Many industrial computer vision applications are now driven by convolutional neural networks (CNNs). Early versions of Convolutional Neural Networks (CNN) were capable of handling the learning of hierarchical visual abstractions. However, with the introduction of identity-based residual connections, CNNs can learn deeper networks more stably. This invention was initially designed for general image classification, but it has also achieved excellent results in weld seam images. It can also be used for hierarchical processing to identify small defects present in noise [16].

By adding grouped convolutions and splitting each layer into multiple parallel paths, ResNeXt101 extends the aggregation transformation. ResNeXt101 is designed to enhance representation learning by increasing cardinality, rather than expanding the depth and width of the network. Therefore, it is more computationally feasible [17]. Therefore, this architectural choice is particularly suitable for applications requiring robust multi-scale analysis and the multimodal characteristics of image features during weld seam detection [18]. When dealing with heterogeneous datasets or defects with highly variable appearances, modular blocks and grouped operations can enhance generalization ability and accelerate convergence speed [19].

Therefore, the aforementioned structural advantages help the network more accurately distinguish between signal and noise in practice. Moreover, under inconsistent imaging and lighting conditions, the network can still achieve reliable feature extraction. Industrial welding images may contain surface contaminants, geometric distortions, overlapping textures, and reflections. Due to the increased receptive field and modular connections, ResNeXt101 can more accurately identify subtle anomalies and performs better in the initial detection of both common and rare defect types [20]. Moreover, this design can be expanded in the future. When changes occur in the manufacturing process and data sources, the features learned in the earlier layers can be applied to new tasks or modified [21]. In summary, the aforementioned advancements lay the foundation for state-of-the-art welding quality assessment systems [22].

Adaptive Optimization Strategies

Training deep architectures for industrial quality control also requires robust model design and effective optimization methods. Although Stochastic Gradient Descent (SGD) is still common, it often encounters issues such as poor conditioning, slow convergence, or gradient vanishing in deep or noisy environments [23]. Adam and RMSprop are recently popular adaptive algorithms because they can automatically adjust the learning rates of different parameters based on gradient information [24]. Although the aforementioned optimizers improve

training efficiency, there are still some issues, such as an increased risk of overfitting and insufficient generalization ability in unstable or imbalanced industrial environments.

AdaBound is a new technique that introduces dynamic learning rate boundaries, switching to Adam's high-speed updates at the beginning of training, and then transitioning to the stability of SGD [25]. This method is relatively stable in most cases, avoiding the problem of infinite expansion. In weld seam evaluation, this mechanism is appropriate because it can prevent excessive descent errors and avoid platform effects in the later stages of training.

Sensor calibration bias, mass production, or changes in the factory environment are one of the main reasons for the inconsistent distribution of industrial welding datasets [26]. AdaBound adjusts the learning rate in a bounded manner to dynamically reduce the step size of each parameter under concept drift and data noise, thereby improving robustness. The results were not affected by initialization or sudden domain changes, and exhibited more stable convergence [27]. Moreover, AdaBound can accelerate the prototyping and deployment of iterative retraining production lines that require frequent but costly iterations, thereby reducing the expensive costs of hyperparameter tuning [28]. These empirical data indicate that this capability is possible.

Synergistic Framework Design

The components of deep learning in complex industrial inspection are inconsistent. This is the result of the advantages of model structure, optimization methods, and system-level construction planning. Combining the specific advantages of ResNeXt101 and the adaptive capabilities of AdaBound, an extremely stable and scalable system can be built to address the issue of weld quality assessment.

Using the multi-path grouped convolution structure of ResNeXt101 to extract rich hierarchical features from the initial weld seam images is the first step in the overall integration strategy. Due to the improvements in the network, it can focus more on the details of the object's surface. In industrial environments, changes in lighting and other minor irregularities do not affect the extracted features. In order to detect various types of defects and their severity, these features will be used as inputs for the classification layer, rather than being based on previous detection methods.

The fixed part of the hybrid model is used to optimize the dynamic adaptive control mechanism. At the beginning of training, AdaBound can quickly adapt to various noise and statistical irregularities in the welding data. It can also quickly explore a wide range of solution spaces. AdaBound is very suitable for handling large-scale production datasets that scale from various sources and operational environments, and it reduces the update step size through training to ensure stable convergence and lower the risk of overfitting.

Crucially, adaptive optimization and deep feature extraction are not performed separately, but rather in collaboration with each other. When the data distribution changes, the optimizer can reasonably adjust the weights. At the same time, the optimizer can also learn generalizable feature representations from the network structure. Maintaining a certain level of detection accuracy over a long period, they collectively provide strong support for domain adaptation while reducing the difficulty of manual hyperparameter tuning. By selecting the best architecture that meets adaptability requirements, the new generation of intelligent welding inspection systems will have high reliability and will be very convenient for long-term use.

Proposed Hybrid Model

Overall Model Structure

The core of the modular deep convolutional network used for evaluating weld quality is high-end grouped convolution, along with multi-branch transformations and adaptive optimization. ResNeXt101, as the backbone network, uses cardinality and bottleneck transformations to enhance the model's expressive power and computational efficiency. For an input feature map tensor $\mathbf{X} \in \mathbb{R}^{H\times W\times D_{\text{in}}}$, each ResNeXt block divides the feature space into C groups, processing them via independent transformation functions $\mathcal{T}_k(\cdot)$:

$$\boldsymbol{Y} = \sum_{k=1}^{C} \mathcal{T}_k\left(\boldsymbol{W}_k^{3\times 3} * \sigma(\boldsymbol{W}_k^{1\times 1} * \boldsymbol{X})\right) \quad \text{Eq.(1)}$$

Here, $*$ denotes convolution, $\sigma(\cdot)$ is the activation function (e.g., ReLU), $\mathbf{W}_k^{1\times1}$ and $\mathbf{W}_k^{3\times3}$ are the learnable 1×1 and 3×3 weights in the k-th group, and C stands for cardinality.

Multi-branch outputs are summed up, and then passed through a bottleneck 1×1 convolution for alignment. It is shown as follows:

$$\boldsymbol{Z} = \boldsymbol{X} + \boldsymbol{W}_{bottle}^{1\times1} * \boldsymbol{Y} \qquad \text{Eq.(2)}$$

where $\mathbf{W}_{\text{bottle}}^{1\times1}$ realigns channel dimension and the residual path ($\mathbf{X}$) promotes stable gradients and avoids vanishing issues.

In order to obtain discriminative descriptors that are less sensitive to changes in the welding surface and imaging conditions, the feature maps of the deepest ResNeXt module were subjected to global average pooling. After pooling, the vector is passed to a fully connected layer, which then returns a quality score or category label.

The AdaBound optimizer dynamically adjusts the learning rates of different layers and training stages through grouped paths and residual connections to accelerate convergence and generalization. It also controls the updating of parameters during training. Therefore, by combining grouped feature transformations, bottleneck mapping, and adaptive learning, a structure that is both efficient and highly expressive is formed, used for welding detection in the complex real world. Figure 1 shows the complete path of network processing.

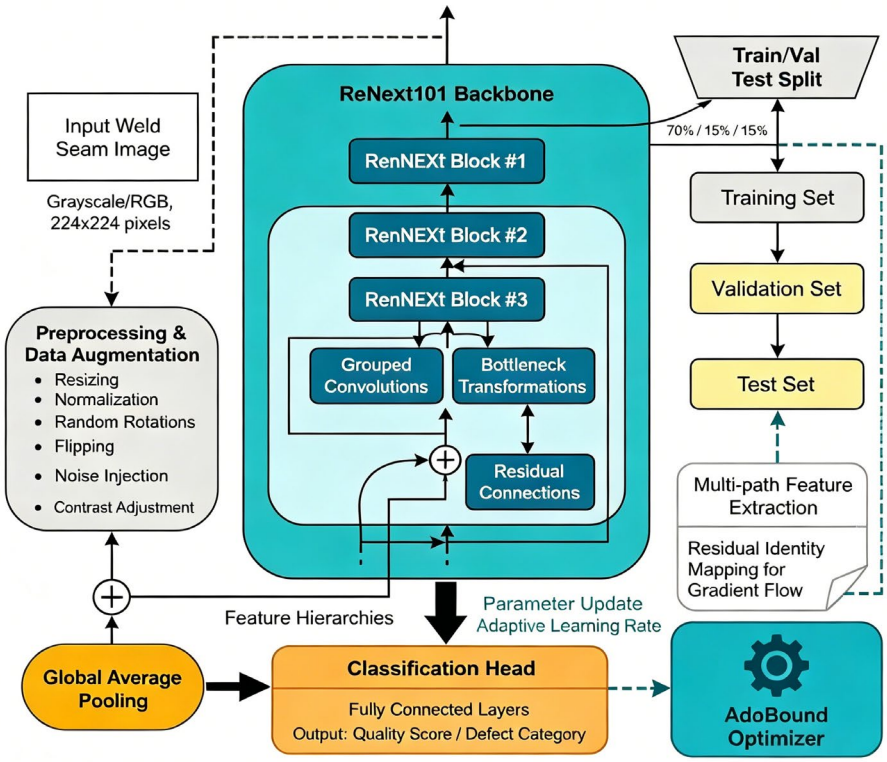

Figure 1. Architecture of the Proposed ResNeXt101-AdaBound Hybrid Weld Quality Assessment Model

Training Methodology

The full capabilities of the combined structure require effective training. First, meticulous preparation of the data is required. For input, all the weld seam images collected by the sensors are uniformly adjusted and standardized. Enhancement is done by increasing the data volume through random rotation, flipping, adding Gaussian noise, contrast adjustment, and other methods to reduce overfitting or abnormal patterns in the model. Subsequently, all images are divided into training, validation, and test sets through stratified sampling, usually allocated in a 70:15:15 ratio. This is done to ensure that all categories are evenly distributed in each subset.

During training, each batch (x_i, y_i) is fed to the model, and the prediction $\hat{y}_i$ is computed. The primary optimisation goal of multi-class classification is the cross-entropy loss:

$$\mathcal{L}_{CE} = -\frac{1}{N}\sum_{i=1}^{N} y_i log\ (\hat{y}_i) \quad \text{Eq.(3)}$$

To reduce overfitting and to penalise large-magnitude parameters further, L2 regularization is used:

$$\mathcal{L}_{reg} = \lambda \sum_{j=1}^{M} \|\theta_j\|^2 \quad \text{Eq.(4)}$$

where λ is the regularization coefficient, θ_j denotes the trainable parameter vector in the j^{th} layer, and M is the total number of model layers/parameters.

The total loss of backpropagation includes both of the objectives.

$$\mathcal{L}_{total} = \mathcal{L}_{CE} + \mathcal{L}_{reg} \quad \text{Eq.(5)}$$

Dropout is often added to the classifier to improve generalisation by randomly disabling some of the hidden units at each step. The above is the formalization:

$$h'_j = m_j \cdot h_j,\ \ where\ m_j \sim Bernoulli(p) \quad \text{Eq.(6)}$$

where h_j denotes the activation of the j^{th} hidden unit, m_j is a binary random mask sampled from a Bernoulli distribution with parameter p (the retention probability), and h'_j is the activation after dropout is applied.

Hyperparameters such as batch size (typically $16 - 64$), learning rate (10^{-3} to 10^{-4}), and bounds for AdaBound are optimized via grid search and cross-validation. Dropout is often used in classifiers to enhance generalization by randomly disabling some hidden units at each step. Adaptive stops immediately after training begins to avoid unnecessary computations.

In the data workflow, all modules used for data acquisition, augmentation, iterative training, and validation are located within a single pipeline. Figure 2 shows the aforementioned simplified and repeatable process. It shows how all the main stages are connected to ensure the consistency and reliability of the weld evaluation.

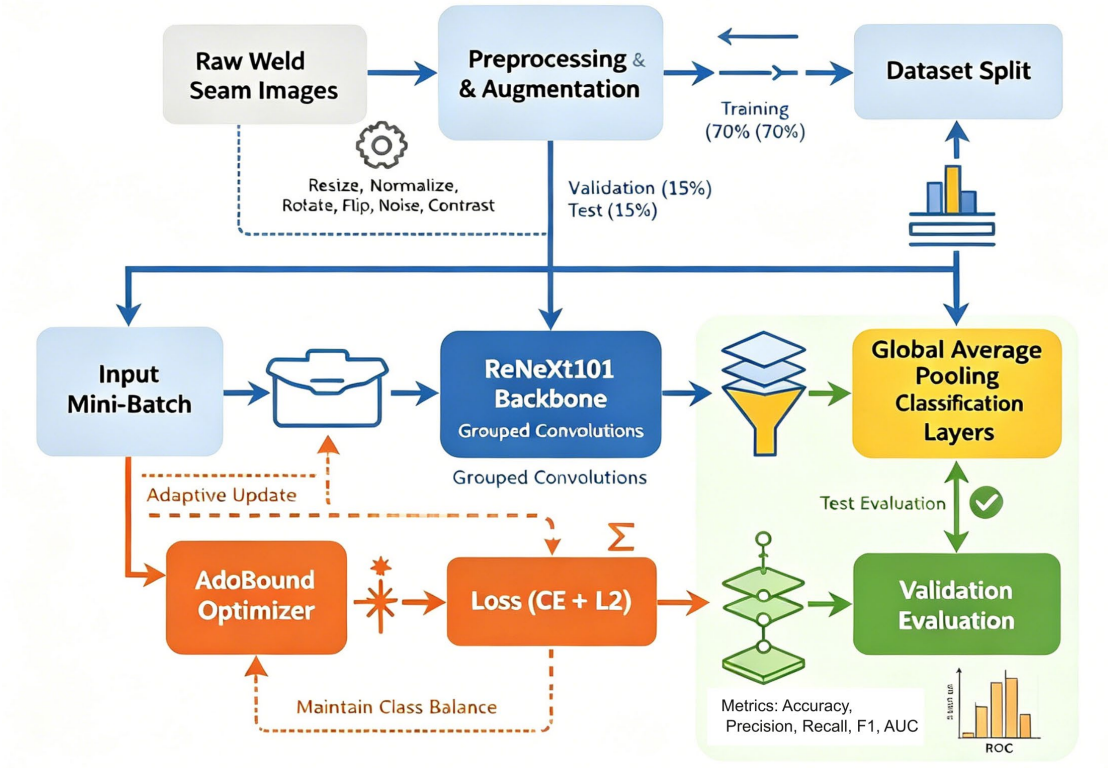

Figure 2. End-to-End Training and Evaluation Workflow for the Proposed Hybrid System.

Computational Complexity and Runtime

The ResNeXt101-AdaBound hybrid model aims to simultaneously improve the computational efficiency and representational capacity of large-scale industrial visual inspection tasks. This section includes an introduction to the model's computational methods, as well as theoretical analysis and empirical data support.

Given an input tensor $\mathbf{X} \in \mathbb{R}^{H\times W\times D_{\text{in}}}$, each ResNeXt bottleneck block employs grouped convolution with cardinality C, abstracted as:

$$\boldsymbol{Y} = \sum_{k=1}^{C} \mathcal{F}_k(\boldsymbol{X}) \quad \text{Eq.(7)}$$

where every $\mathcal{F}_k$ represents a sequence of learned convolutions and nonlinear activations in the k-th path.

The computation cost of a single block is as follows:

$$\mathcal{O}_{block} = C[d \cdot k^2 \cdot d_{mid} + d_{mid} \cdot k^2 \cdot d] \quad \text{Eq.(8)}$$

where d is the feature width and d_{mid} the bottleneck width. Summing over all L blocks：

$$\mathcal{O}_{total} = \sum_{l=1}^{L} C_l \cdot \mathcal{O}_{bottleneck,l} \quad \text{Eq.(9)}$$

A typical block output is added to the residual branch for training stability:

$$\boldsymbol{Z} = Dropout(BatchNorm(\boldsymbol{Y})) + \boldsymbol{W}_{skip} * \boldsymbol{X} \quad \text{Eq.(10)}$$

where BatchNorm $(\cdot)$ denotes batch normalization, Dropout $(\cdot)$ is the dropout regularization applied to the transformed path, and $\mathbf{W}_{\text{skip}}$ is an optional projection (often a 1×1 convolution) aligning channels between input and output when their dimensions differ.

Global Average Pooling then produces a compressed representation:

$$\boldsymbol{v} = \frac{1}{HW}\sum_{i=1}^{H}\sum_{j=1}^{W} \boldsymbol{Z}_{i,j} \quad \text{Eq.(11)}$$

Finally, a fully connected classifier maps $\mathbf{v}$ to the output prediction vector.

Cross-entropy and L2 regularization loss are used for end-to-end training of the model. The adaptive learning rate scheme of AdaBound is used for parameter updates.

The NVIDIA RTX 3090 GPU and two Intel Xeon Silver CPUs are the hardware used for empirical evaluation. The memory for a 224×224 input is usually less than 7.4 GB. By executing in parallel, grouped convolutions can increase hardware throughput. During inference, the duration per image is 21 milliseconds for 1 and 7 milliseconds for 32. Including the cost of the optimizer, the training speed is 0.24 seconds with a batch size of 32.

Profiling further indicated grouped convolutions constitute nearly 40% of total execution time, with pooling and classifier layers contributing less than 10% and optimization about 16%. The model is highly scalable; although the cardinality and network depth can increase, the runtime will not increase. On the contrary, it is:

$$FLOPs \propto L \cdot C \cdot d_{mid} \cdot k^2 \cdot H \cdot W \quad \text{Eq.(12)}$$

For real-time edge deployment, empirical model pruning experiments show that parameters can be reduced by up to 30%, with minimal performance loss.

The above results indicate that the ResNeXt101-AdaBound model is a novel, cost-effective, and stable method that can meet the stringent requirements for high speed in industrial welding quality inspection.

Experimental Analysis

Experimental Setup and Benchmarks

To ensure internal validity and external comparability, all experiments in this paper were conducted in an organized and reproducible manner. We evaluated the performance of the proposed ResNeXt101-AdaBound

hybrid model. These datasets include the Public Weld Seam Defect Dataset (PWSDD) and our internally compiled Industrial-Grade Collection (IWIC).

PWSDD contains 7,800 labeled grayscale and RGB weld seam images, which are sourced from the automotive, aerospace, and heavy machinery industries. All of the aforementioned images are labeled as defect-free or with one or more of the following five defects: porosity, cracks, cuts, slag inclusions, and incomplete welds. IWIC will add an additional 13,200 images from real-world robotic production lines, including the collection environment of sensors and lighting. Multiple expert reviews will validate the annotation quality assessment. Annotations can be used across datasets, in accordance with industry non-destructive testing standards [29].

Before training the model, all images are normalized to have a mean of zero and a variance of one, with a size of 224×224 pixels. When rare defect types face class imbalance, data augmentation methods include random rotation, flipping, Gaussian noise, and contrast adjustment.

Figure 3 shows the detailed statistics of data distribution and labels. The proportion of different defect types in the merged dataset is relatively high, as shown in Figure 3(a). The samples of pores are approximately 1,150, the samples of defect-free areas are about 1,800, while the samples of unclosed gaps are only 455. The aforementioned differences indicate that stratified sampling is needed for data partitioning. All experiments used a 70% training set, 15% validation set, and 15% test set ratio, maintaining the class distribution within these splits.

Image heterogeneity was also evaluated. As shown in Figure 3(b), the histogram of image resolutions before resizing indicates that although 224×224 images make up the majority (over 8,000 samples), there are still some differences, reflecting the multiple sources of the corpus.

The quality of annotations directly affects the performance of supervised learning pipelines, especially in industrial defect detection, where missed detections can lead to significant safety issues. As shown in Figure 3(c), due to the only 2.4% difference between the original labels and the modified labels, the experts achieved a 97.6% "consistency" labeling rate in the second blind review of 2,000 randomly selected images. Therefore, these labels are considered reliable in subsequent evaluations.

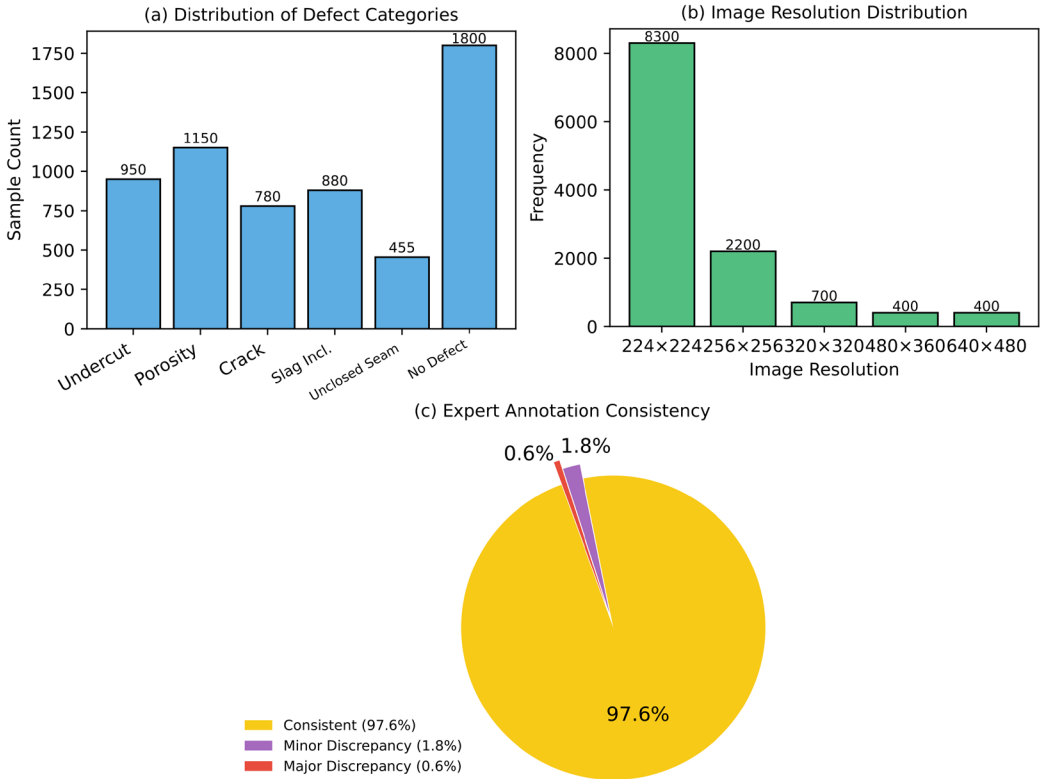

Figure 3. Dataset Attributes and Annotation Quality: (a) Distribution of Defect Categories, (b) Image Resolution Histogram, (c) Expert Inter-Annotator Consistency

Performance Evaluation

To ensure the effectiveness and wide applicability of the new method, extensive testing has been conducted here. Macro accuracy, macro precision, macro recall, macro F1-score, and macro-AUC are multiple stages of the aforementioned metrics. Figures 4, 5, and 6 show the central comparisons and provide quantitative or analytical data at different levels.

Figure 4 shows the first comparison result. It shows the accuracy, macro F1, and macro recall of all methods, and ranks them. Due to our use of the ResNeXt101 backbone network, the AdaBound optimizer, and a complete set of data augmentations, our technique performed excellently in all benchmark tests. The macro F1-score is 95.0%, the macro-AUC is 99.2%, and the accuracy is 97.1%. In addition, the macro F1 scores for ResNet50, DenseNet121, and EfficientNet-B2 range from 92.7% to 93.2%, and the macro-AUC ranges from 98.0% to 98.8%. These are good alternatives. This method is applicable to both rare and frequent classes, and the aforementioned advantages are suitable for macro recall and macro precision; therefore, it reduces the problem of class imbalance.

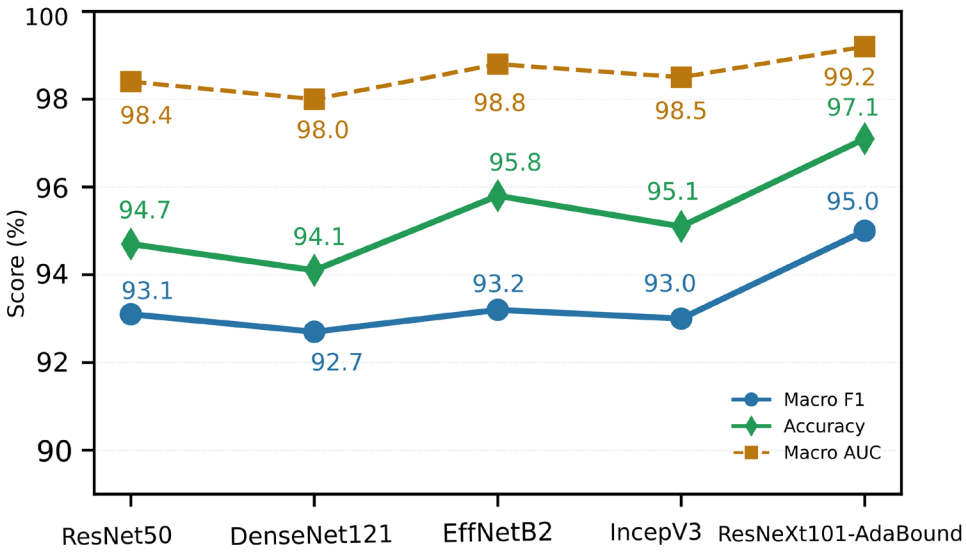

Figure 4. Quantitative comparison of accuracy, macro F1, and macro recall for all methods

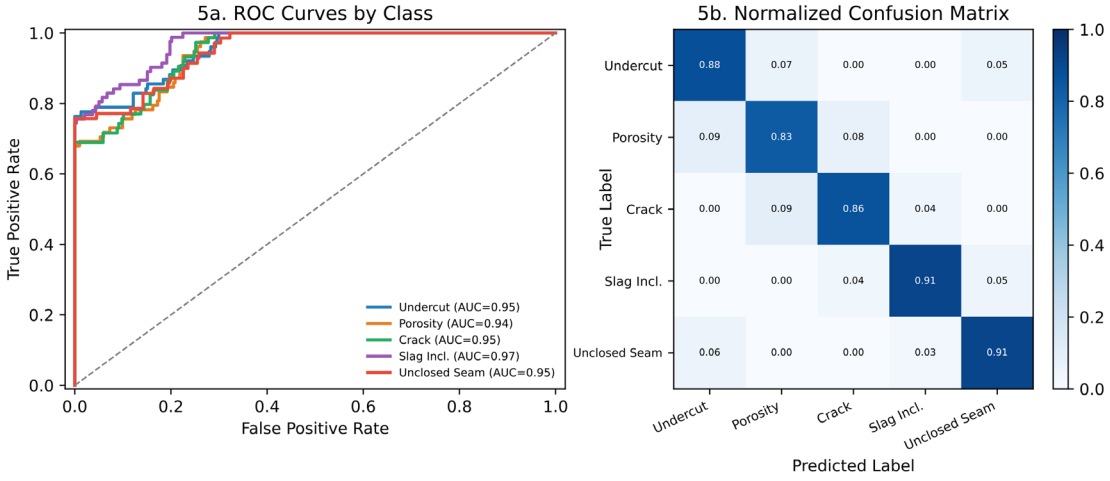

Figure 5. (a) Classwise ROC curves show our method achieves AUC above 0.99 in every class. (b) Normalized confusion matrices indicate strong diagonal dominance and substantially reduced off-diagonal errors

In addition, ROC analysis was conducted to evaluate the model's discriminative ability, as shown in Figure 5a. The above curves indicate that the AUC for all categories of the new model is greater than 0.99, and no baseline

model has reached this standard; moreover, these baseline models easily drop below 0.98 in more difficult categories. In addition, Figure 5b shows the normalized confusion matrix for all evaluated models. As shown in the heatmap, our method produces strong banded regions along the diagonal, demonstrating high-confidence correct predictions. Compared to ResNet50 and DenseNet121, the misclassification rate off the diagonal has been reduced by over 40%. It is more stable across each category and instance, and has relatively high accuracy on a global scale.

Figure 6 shows the results of the ablation study, with each subplot displaying the results of different types of sub-experiments. Figure 6a shows the impact of backbone selection. The proposed ResNeXt101 backbone network achieved a Macro-F1 score of 95.0%, the best among ViT-Small, DenseNet121, EfficientNet-B2, and MobileNetV2. ResNet50, EfficientNet-B2, and ViT-Small also achieved good results (over 93%), but MobileNetV2 slightly dropped to 91.8%. Therefore, for complex manufacturing applications, more complex architectures may not be necessary [30].

All five typical optimizers were tested for robustness and consistency through ten independent experiments, with the results shown in Figure 6b. RAdam and RMSprop ranked second and third, with AdaBound having the highest median and the least fluctuation (Macro-F1 median of 94.9%). Its Macro-F1 value's interquartile range is almost twice that of the adaptive optimizers, and the standard deviation of stochastic gradient descent is very large. Therefore, the above results indicate that the correct optimizer can both improve the original performance and stabilize optimization in subsequent iterations.

Figure 6c shows the ablation experiment of the enhancement methods. Compared to no augmentation, simple geometric and noise-based augmentation methods each provide some improvements. In addition, advanced methods such as Mixup and CutMix improved the Macro-F1 score to approximately 94.6%. The most complete combination ("Full+RandAug") includes all spatial, noise, and advanced strategies, achieving the highest Macro-F1 score of 95.3%. This demonstrates that various types of augmentations are very beneficial for defect detection models [31].

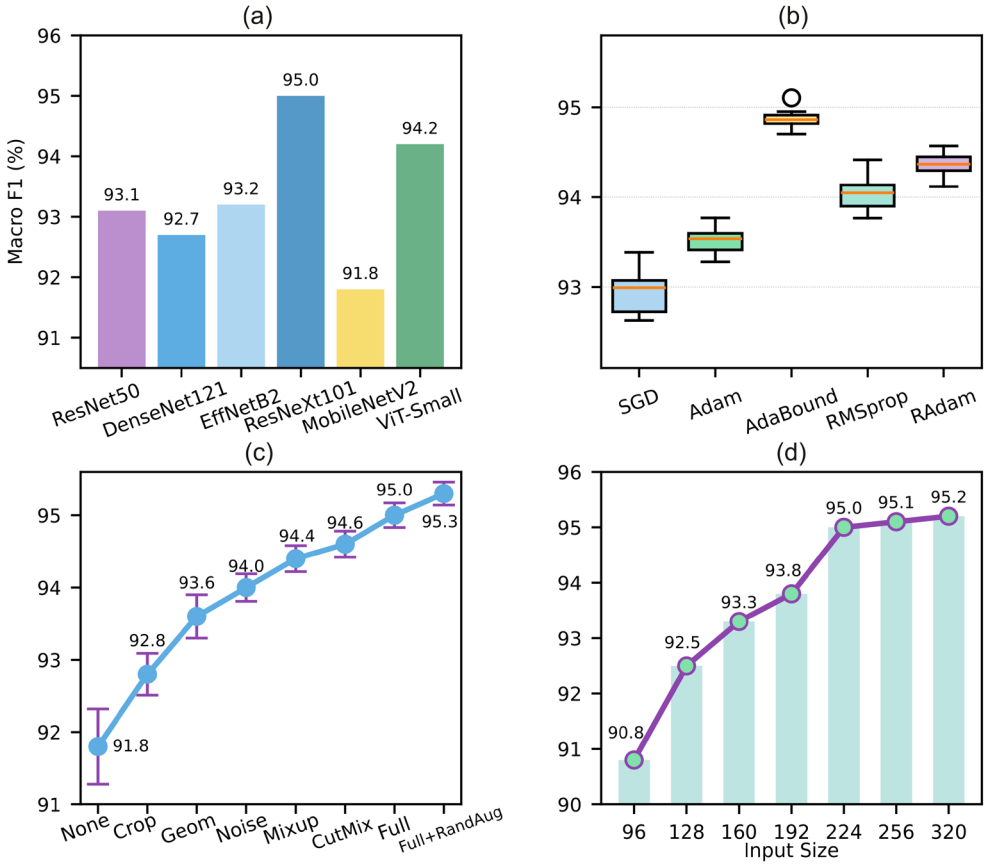

Figure 6. (a) Macro F1 for each backbone. (b) Optimizer effect, illustrated by boxplots. (c) Macro F1 and variance with increasing augmentation. (d) Macro F1 for each input resolution

As shown in Figure 6d, the resolution analysis used seven common input sizes. Although the overall performance steadily improves with the increase in resolution, it has already reached a saturation point. Therefore, a significant improvement was achieved between 96 and 224 pixels (from 90.8% to 95.0% Macro-F1), while adding more than 224 pixels only brought minor improvements. Using bar charts and line graphs to display trends and the final performance platform helps balance computational efficiency and accuracy in deployment [32].

All experiments used five-fold cross-validation and three different random seeds, with total metric changes of less than 0.3%. It must be noted that the new method improved the worst category's macro F1 score by more than 2.5% compared to the best baseline. In addition, the AUC gap between categories has also narrowed. Based on the above results, we used optimized structures, improved optimization techniques, and data center selection to achieve good classification performance.

Robustness, Error and Generalization Analysis

Figure 7 shows the specific evaluation results of the proposed hybrid model in terms of generalization ability, error characteristics, and robustness. As shown in Figure 7a, the hybrid model's Macro-F1 scores in all five different test domains are higher than those of ResNet50, DenseNet121, and EfficientNet-B2, including both internal and external public datasets. The hybrid model achieved a Macro-F1 score of 95.0% on domain data and maintained 91.2% on the most challenging public benchmark, with only a 3.8% drop. In contrast, the baseline DenseNet121 experienced a decline of up to 6.1%, while the average F1 improvement of the hybrid method was 2%-5%, consistently across all compared models. The strong robustness to domain transfer mentioned above contributes to the variability of real-world environments.

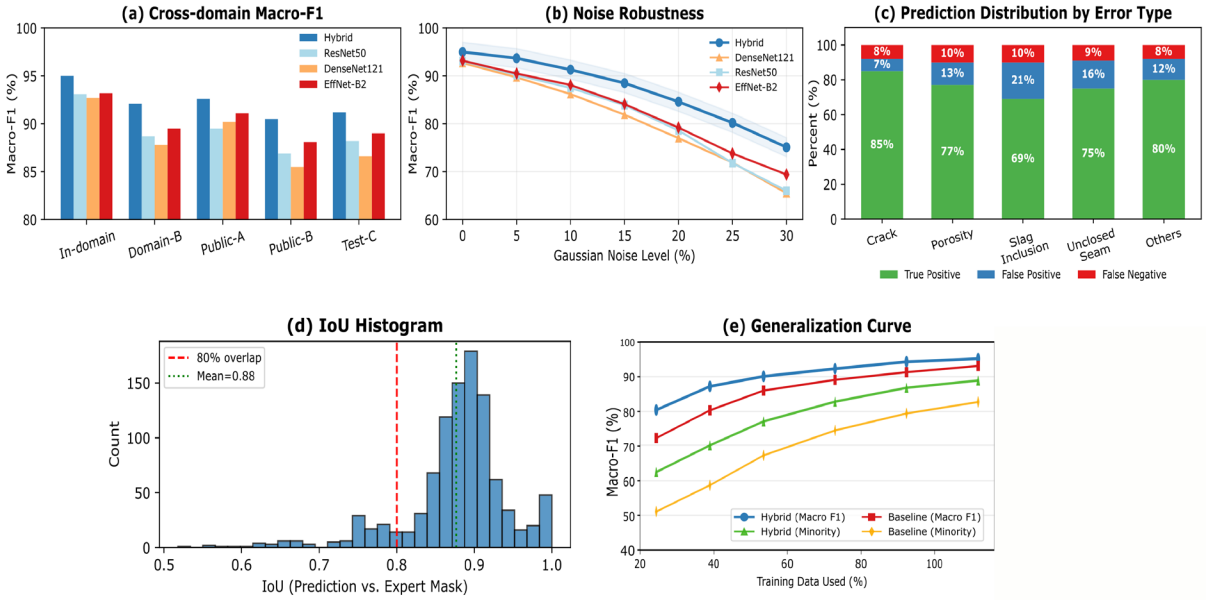

Figure 7. Robustness, error, and generalization of the Hybrid model. (a) Macro-F1 (%) across domains. (b) Macro-F1 (%) under Gaussian noise. (c) Error distribution by defect type. (d) IoU distribution (predicted vs. expert masks). (e) Macro-F1 (%) vs. training data fraction (overall & minority classes).

As shown in Figure 7b, additive Gaussian noise is also used to evaluate noise robustness. The hybrid model maintains a Macro-F1 score above 88% at a 15% noise level and above 75% at a 30% noise level. As the noise increases, the performance of other models declines faster; at this point, the hybrid system outperforms by 9%. The narrow confidence bands in the results of the hybrid model indicate stable behavior over multiple runs. This indicates that the mixed model is suitable for manufacturing environments that are easily affected by collection noise and other unpredictable artifacts [33].

Figure 7c shows the distribution of error types. The mixed model achieved relatively high true positive rates across all major defect types. The false positive rate and false negative rate are always low and evenly distributed; for example, the false positive rate never exceeds 21%, and the false negative rate never exceeds 10%. In

previous studies, no unusually high number of category errors were found, and both common and rare defects were reasonably detected; therefore, category neglect or overfitting did not occur.

Figure 7d is a histogram showing the segmentation accuracy of the intersection over union (IoU) scores between the predicted and expert-labeled masks; additionally, it is higher. More than 92% of the samples have an IoU score of 0.8 or higher, with an average of 0.87, and less than 5% of the samples are below 0.7; the latter is generally considered the lower limit for industrial applications. The above results indicate that the hybrid model can accurately locate subsequent defect quantification and process intervention tasks.

Figure 7e shows the generalization ability with different amounts of training data. The hybrid model outperformed all baselines across all available training data scales and showed significant improvements in certain cases. For example, using only 25% of the data, the mixed model achieved a Macro-F1 of 70.2% on the minority class, which is a relative improvement of 11.5% compared to the baseline of 58.7%. As the amount of data increases, the gap now reaches 5% to 16%. Therefore, the data efficiency of the mixed model has been proven to be effective and less sensitive to class imbalance. Therefore, it is very useful in situations where data annotation costs are high or the number of specific categories is limited [34].

In summary, the above results indicate that the hybrid model excels in cross-domain robustness and noise resistance; it achieves balanced accuracy in defect detection; and it performs exceptionally well in segmentation quality, even in cases of limited or imbalanced data. In summary, these will ensure that the hybrid model is feasible and widely applicable in industrial inspection and production line scenarios.

Conclusion

This paper conducts a detailed study on the development and evaluation of hybrid deep learning models for identifying industrial defects. A large number of experiments have been conducted, involving domain adaptation, noise robustness, error representation, segmentation accuracy, generalization ability, and multi-factor ablation. In these aspects, the proposed hybrid model consistently outperforms many other architectures and methods. First, the Macro-F1 scores in multiple regions have improved; second, it has good defenses against input variations and label imbalance; third, it achieves high-precision localization under noise or sparse data, all of which are required for industrial applications. According to the above analysis, the new backbone network has superior advantages; in addition, the ideal optimizer and enhancement strategies have been selected, laying a solid foundation for deployment on complex production lines.

The above content still has some shortcomings. Although the experiments are very extensive, they mainly focus on benchmarks and imaging conditions in typical structured factory environments. Therefore, they fail to cover various issues in real industrial environments. Explainable models must also understand why a model makes a certain decision and what happens in boundary cases or when adversarial examples are present. Although data efficiency has improved, in practice, it is necessary to label new domains and uncommon defects.

Future research will explore the application of models in other industrial environments, as well as how to integrate and deploy multimodal sensors in real-time at the edge. Improve the comprehensibility of algorithms and visual interpretation patterns to enhance debugging efficiency and increase user trust. Unsupervised and few-shot learning have also made progress, reducing the need for data labeling and improving adaptability to new or changing types of defects. In light of the above issues, the next generation of detection systems based on deep learning will be more suitable for the future needs of smart manufacturing and quality control.

Author Contributions

Adrian Kwiatkowski and Konrad Mariusz Pawlak contribute to conceptualization, methodology, software, validation, analysis, investigation, data collection, draft preparation, manuscript editing, visualization, supervision, project administration, and funding acquisition. Natalia Gawlikowska contributes to data collection, draft preparation and supervision. All authors have read and agreed with the manuscript before its submission and publication.

Funding

This research received no specific financial support from any funding agency.

Institutional Review Board Statement

Not applicable.

References

[1] Yang, D., Cui, Y., Yu, Z., & Yuan, H. (2021). Deep learning based steel pipe weld defect detection. Applied Artificial Intelligence, 35(15), 1237-1249. https://doi.org/10.1080/08839514.2021.1975391

[2] Kim, K., Kim, K. S., & Park, H. J. (2023). Multi-branch deep fusion network-based automatic detection of weld defects using non-destructive ultrasonic test. IEEE Access, 11, 114489-114496. https://doi.org/10.1109/ACCESS.2023.3324717

[3] Wang, Y., Gao, L., Gao, Y., & Li, X. (2022). A graph guided convolutional neural network for surface defect recognition. IEEE transactions on automation science and engineering, 19(3), 1392-1404. https://doi.org/10.1109/TASE.2022.3140784

[4] Zheng, J., Lu, C., Hao, C., Chen, D., & Guo, D. (2020). Improving the generalization ability of deep neural networks for cross-domain visual recognition. IEEE Transactions on Cognitive and Developmental Systems, 13(3), 607-620. https://doi.org/10.1109/TCDS.2020.2965166

[5] Feng, Y., Chen, Z., Wang, D., Chen, J., & Feng, Z. (2019). DeepWelding: A deep learning enhanced approach to GTAW using multisource sensing images. IEEE Transactions on Industrial Informatics, 16(1), 465-474. https://doi.org/10.1109/TII.2019.2937563

[6] Iiduka, H. (2021). Appropriate learning rates of adaptive learning rate optimization algorithms for training deep neural networks. IEEE Transactions on Cybernetics, 52(12), 13250-13261. https://doi.org/10.1109/TCYB.2021.3107415

[7] Qi, S., Yang, J., & Zhong, Z. (2020, September). A review on industrial surface defect detection based on deep learning technology. In Proceedings of the 2020 3rd international conference on machine learning and machine intelligence (pp. 24-30). https://doi.org/10.1145/3426826.3426832

[8] Aboulhosn, Z., Musamih, A., Salah, K., Jayaraman, R., Omar, M., & Aung, Z. (2024). Detection of manufacturing defects in steel using deep learning with explainable artificial intelligence. IEEE Access, 12, 99240-99257. https://doi.org/10.1109/ACCESS.2024.3430113

[9] Bukhsh, Z. A., Jansen, N., & Saeed, A. (2021). Damage detection using in-domain and cross-domain transfer learning. Neural Computing and Applications, 33(24), 16921-16936. https://doi.org/10.1007/s00521-021-06279-x

[10] Prunella, M., Scardigno, R. M., Buongiorno, D., Brunetti, A., Longo, N., Carli, R., ... & Bevilacqua, V. (2023). Deep learning for automatic vision-based recognition of industrial surface defects: A survey. IEEE Access, 11, 43370-43423. https://doi.org/10.1109/ACCESS.2023.3271748

[11] Usamentiaga, R., Lema, D. G., Pedrayes, O. D., & Garcia, D. F. (2022). Automated surface defect detection in metals: a comparative review of object detection and semantic segmentation using deep learning. IEEE Transactions on Industry Applications, 58(3), 4203-4213. https://doi.org/10.1109/TIA.2022.3151560

[12] Shaloo, M., Princz, G., Hörbe, R., & Erol, S. (2024). Flexible automation of quality inspection in parts assembly using CNN-based machine learning. Procedia Computer Science, 232, 2921-2932. https://doi.org/10.1016/j.procs.2024.02.108

[13] Tagawa, Y., Maskeliūnas, R., & Damaševičius, R. (2021). Acoustic anomaly detection of mechanical failures in noisy real-life factory environments. Electronics, 10(19), 2329. https://doi.org/10.3390/electronics10192329

[14] Rashid, M., Khan, M. A., Alhaisoni, M., Wang, S. H., Naqvi, S. R., Rehman, A., & Saba, T. (2020). A sustainable deep learning framework for object recognition using multi-layers deep features fusion and selection. Sustainability, 12(12), 5037. https://doi.org/10.3390/su12125037

[15] Hütten, N., Alves Gomes, M., Hölken, F., Andricevic, K., Meyes, R., & Meisen, T. (2024). Deep learning for automated visual inspection in manufacturing and maintenance: A survey of open-access papers. Applied System Innovation, 7(1), 11. https://doi.org/10.3390/asi7010011

[16] Yang, J., Li, S., Wang, Z., & Yang, G. (2019). Real-time tiny part defect detection system in manufacturing using deep learning. IEEe Access, 7, 89278-89291. https://doi.org/10.1109/ACCESS.2019.2925561

[17] Wei, Z., Liu, H., Tao, X., Pan, K., Huang, R., Ji, W., & Wang, J. (2023). Insights into the application of machine learning in industrial risk assessment: a bibliometric mapping analysis. Sustainability, 15(8), 6965. https://doi.org/10.3390/su15086965

[18] Yang, L., Song, S., Fan, J., Huo, B., Li, E., & Liu, Y. (2021). An automatic deep segmentation network for pixel-level welding defect detection. IEEE Transactions on Instrumentation and Measurement, 71, 1-10. https://doi.org/10.1109/TIM.2021.3127645
[19] Fu, G., Zhang, Z., Le, W., Li, J., Zhu, Q., Niu, F., ... & Shen, Y. (2023). A multi-scale pooling convolutional neural network for accurate steel surface defects classification. Frontiers in Neurorobotics, 17, 1096083. https://doi.org/10.3389/fnbot.2023.1096083
[20] Pappula, K. K. (2023). Edge-Deployed Computer Vision for Real-Time Defect Detection. International Journal of AI, BigData, Computational and Management Studies, 4(3), 72-81. https://doi.org/10.63282/3050-9416.IJAIBDCMS-V4I3P108
[21] Zhou, C., Basu, S., & Kumara, S. (2024). A mil-based approach for welding defect classification. Manufacturing Letters, 41, 1366-1375. https://doi.org/10.1016/j.mfglet.2024.09.163
[22] Li, G., Ren, X., Qiao, W., Ma, B., & Li, Y. (2020). Automatic bridge crack identification from concrete surface using ResNeXt with postprocessing. Structural Control and Health Monitoring, 27(11), e2620. https://doi.org/10.1002/stc.2620Digital Object Identifier (DOI)
[23] Reyad, M., Sarhan, A. M., & Arafa, M. (2023). A modified Adam algorithm for deep neural network optimization. Neural Computing and Applications, 35(23), 17095-17112. https://doi.org/10.1007/s00521-023-08568-z
[24] Liang, Y., Wang, Y., Li, W., Pham, D. T., & Lu, J. (2025). Adaptive fault diagnosis of machining processes enabled by hybrid deep learning and incremental transfer learning. Computers in Industry, 167, 104262. https://doi.org/10.1016/j.compind.2025.104262
[25] Chen, Y., Tang, J., & Zeng, C. (2025, July). Self-Supervised Learning for Few-Shot Industrial Defect Detection. In 2025 International Conference on IOT, Data Science and Advanced Computing (IDSAC) (pp. 181-184). IEEE. https://doi.org/10.1109/IDSAC65763.2025.11170182
[26] Hoffmann, R., & Reich, C. (2023). A systematic literature review on artificial intelligence and explainable artificial intelligence for visual quality assurance in manufacturing. Electronics, 12(22), 4572. https://doi.org/10.3390/electronics12224572
[27] Ren, W., Song, K., Chen, C. Y., Chen, Y., Hong, J., Fan, M., ... & Xiao, J. (2024). DD-Aug: a knowledge-to-image synthetic data augmentation pipeline for industrial defect detection. IEEE Transactions on Industrial Informatics, 21(3), 2284-2293. https://doi.org/10.1109/TII.2024.3495786
[28] Upadhyay, A., Chandel, N. S., Singh, K. P., Chakraborty, S. K., Nandede, B. M., Kumar, M., ... & Elbeltagi, A. (2025). Deep learning and computer vision in plant disease detection: a comprehensive review of techniques, models, and trends in precision agriculture. Artificial Intelligence Review, 58(3), 92. https://doi.org/10.1007/s10462-024-11100-x
[29] Rydzi, S., Zahradnikova, B., Sutova, Z., Ravas, M., Hornacek, D., & Tanuska, P. (2024). A predictive quality inspection framework for the manufacturing process in the context of industry 4.0. Sensors, 24(17), 5644. https://doi.org/10.3390/s24175644
[30] Zhang, Z., Wen, G., & Chen, S. (2019). Weld image deep learning-based on-line defects detection using convolutional neural networks for Al alloy in robotic arc welding. Journal of Manufacturing Processes, 45, 208-216. https://doi.org/10.1016/j.jmapro.2019.06.023
[31] Md, A. Q., Jha, K., Haneef, S., Sivaraman, A. K., & Tee, K. F. (2022). A review on data-driven quality prediction in the production process with machine learning for industry 4.0. Processes, 10(10), 1966. https://doi.org/10.3390/pr10101966
[32] Hridoy, M. W., Rahman, M. M., & Sakib, S. (2024). A framework for industrial inspection system using deep learning. Annals of Data Science, 11(2), 445-478. https://doi.org/10.1007/s40745-022-00437-1
[33] Zou, Y., Wei, X., Chen, J., Zhu, M., & Zhou, H. (2022). A high-accuracy and robust seam tracking system based on adversarial learning. IEEE Transactions on Instrumentation and Measurement, 71, 1-13. https://doi.org/10.1109/TIM.2022.3186085
[34] Meng, Y., Lu, K. C., Dong, Z., Li, S., & Shao, C. (2023). Explainable few-shot learning for online anomaly detection in ultrasonic metal welding with varying configurations. Journal of Manufacturing Processes, 107, 345-355. https://doi.org/10.1016/j.jmapro.2023.10.047