Data Engineering and Applications

Research Paper

Real-Time Surface Quality Assessment Using Patch-level Attention ResNet

Magdalena Król[1] and Michał Tomasz Dąbrowski[1, *]

[1] Faculty of Computer Science, Adam Mickiewicz University, 61-614, Poznań, Poland
*Corresponding author: michal.d@amu.edu.pl

Abstract. In recent years, surface quality inspection based on computer vision has frequently been used in high-precision, reliable automated manufacturing to identify defects. To address subtle, irregular, and low-contrast surface defects, this paper proposes a real-time surface quality assessment framework based on a patch-level attention-enhanced ResNet architecture. Using adaptive patch division and two different types of attention mechanisms, this new technique simultaneously captures local features and global relationships, and it is interpretable. In order to handle different environments, the preprocessing module in this paper will apply contrast limitation, random augmentation, and adaptive histogram equalization. The convolutional feature extractor divides each input image into multiple patches. Then, the channel and spatial attention modules recalibrate the feature maps based on defect relevance. Subsequently, the recalibrated features are integrated into the regression head for continuous surface quality score prediction and the classification head for defect category identification. Experiments have been conducted using large-scale public and private industrial datasets containing over 40,000 labeled images. The above results indicate that the system has an inference speed of over 38 frames per second and an average precision mean of 0.946. Therefore, it surpasses both the baseline and other top solutions in terms of accuracy and speed. Ablation studies indicate that channel attention and spatial attention are complementary, and their combination still outperforms single-path designs. The all-encompassing framework for autonomous surface inspection in manufacturing has high reliability, is easy to interpret, and performs well.

Keywords: *Image Processing, Surface Defect Detection, Patch-Level Attention, Hybrid Attention Mechanism, Real-Time Quality Assessment*

Received on 20 June 2025, Accepted on 17 December 2025, Published on 05 January 2026

Introduction

In order to improve production efficiency, intelligent surface quality inspection systems have been widely used for quality control and automated manufacturing [1]. Due to the higher demands of the aerospace, electronics, and automotive industries for high-value, defect-free parts, reliable and accurate surface inspection systems have been introduced [2]. Previous inspection methods used handcrafted features and subjective visual assessments, making them labor-intensive and often unable to detect minor defects or adapt to various types of products [3]. Due to the aforementioned shortcomings of traditional computer vision techniques, many researchers are focusing on deep learning and other highly representative models [4]. The aforementioned methods are generally easier to classify and locate than rule-based systems [5]. Most existing deep models are black boxes and are not interpretable; therefore, they cannot be applied to new surface patterns [6]. Moreover, due to irregular shapes, low contrast, and diverse backgrounds, the detection framework must be both precise and highly adaptable [7]. Therefore, comprehensive and accurate surface assessment remains a current focus [8].

Although some progress has been made recently, many issues still hinder the widespread use of general surface assessment networks in industry [9]. Training deep learning models usually requires a large amount of labeled data. However, industrial data is often imbalanced, domain-specific, and expensive to label [10]. The structural

and appearance differences of surface defects are significant, ranging from minor scratches to severe deformations. Therefore, feature extraction is very difficult and often leads to missed detections or false detections of out-of-distribution samples [11]. New attention mechanisms have recently been proposed to help networks focus on important areas. However, when they are combined with other methods, some design and optimization issues arise, particularly in balancing local and global defect representations [12]. Moreover, there are issues with the model's robustness, inference speed, and capacity, and real-time performance and edge deployment are limited [13]. The issues of high-resolution analysis, parameter efficiency, and interpretability in industrial inspection processes have not yet been fully resolved, despite the widespread use of advanced technologies such as transformers and hybrid CNN models [14]. The aforementioned unresolved issues indicate that the surface evaluation capabilities based on deep learning are currently insufficient to meet the needs of the manufacturing industry [15]. Regarding accuracy, adaptability, and interpretability, there is currently a lack of a complete framework [16].

To address the aforementioned issues, we propose an all-weather surface assessment system. The system has reliability, is easy to operate, and has efficient defect localization capabilities, making it suitable for use in various industries. In order to dynamically handle local defects and global changes on the surface, we adopted a feature aggregation scheme based on attention mechanisms, which also supports adaptive partitioning. In recent years, many researchers have studied methods to improve the interpretability of computer vision systems while enhancing detection accuracy and attention mechanisms. Further provide ablation studies, including the design of attention modules, and quantitatively and intuitively evaluate their impact on model interpretability and accuracy. A large number of experiments were conducted on open-source and proprietary industrial datasets, and the proposed system consistently outperformed previous best solutions in both quantitative metrics and visual surface evaluations. The system is scalable and practically applicable.

Related Work

Surface Defect Inspection Approaches

Automated quality assurance in manufacturing has always been the foundation of surface defect detection [17]. Early research in this field primarily relied on traditional computer vision processes, such as manually creating edge contours, texture descriptors, and grayscale histograms to extract defect metrics from surface images [18]. These methods perform excellently in controlled environments, but they are less stable under conditions of lighting changes, complex backgrounds, or objects made of different materials [19]. In order to improve generalization ability, various advanced signal processing techniques, such as adaptive filtering and wavelet analysis, are now being used. However, these techniques are still insufficient in terms of parameter settings and widespread application across all product lines, and they are rarely used across all product lines as well [20]. Rule-based heuristic methods and template matching are also frequently used, but these solutions are often not suitable for defects, deformations, or occlusions of various sizes that occur in practice [21]. Although there are fewer new defect pattern categories, handcrafted features have been used to identify defects.

Due to the drawbacks of manual methods, people have started using data-driven machine learning classifiers, such as ensemble models and support vector machines, and employing statistical features to enhance discriminative power [22]. However, these models based on traditional learning also rely on the quality, diversity, and relevance of features. Therefore, they are difficult to perform well in complex or new environments. As the complexity of industrial products and the precision required for defect detection increase, it has become evident that the aforementioned methods cannot achieve accurate detection. A more robust and adaptable detection framework is needed, as there are still issues with false positives, missed subtle defects, and difficulties in handling high intra-class variance [23]. Therefore, the field is now beginning to use automated deep representation learning.

Deep Learning and Attention Mechanisms

The new method for surface inspection is based on deep learning [24]. Convolutional Neural Networks (CNNs) excel at automatically learning hierarchical features, which makes them more robust to noise and image variations in defect detection [25]. In order to achieve high accuracy with low deployment costs, many current surface inspection solution architectures are based on deep residual networks or lightweight mobile models.

Since they are used for end-to-end learning, manual feature engineering is not required. In addition, they perform well in many industrial applications [26].

Systems based on traditional CNNs still cannot identify fine-grained defects with low contrast or sparse distribution, despite some improvements. The attention mechanism is introduced in visual models to allow the network to focus on relevant parts of the image or feature map. To improve the model's detection accuracy and interpretability, channel and spatial attention modules can recalibrate feature maps. The Transformer-based architecture has recently made significant progress in global self-attention. It is also used for visual anomaly detection to model long-range dependencies and complex relational patterns, which are difficult for traditional CNN models to learn. The addition of attention mechanisms can improve the model's performance and help interpret the model; both are necessary for regulations and large-scale industrial applications. Nevertheless, the task of creating an integrated system that can simultaneously enhance the recognition of three dimensions is still not accomplished. At the current stage of development in surface defect assessment research, we still hope to find effective solutions to address these issues.

Framework Design

Dataset and Preprocessing

Conduct experiments using private industrial samples and popular open-source surface defect datasets, according to the aforementioned surface evaluation framework. Open-source benchmarks such as the DAGM series and the Severstal Steel dataset record a large number of materials and various defects. At the same time, private datasets are used to ensure the stability of this method in practical industrial applications.

In order to ensure network compatibility, each image undergoes regular quality checks and is standardized to the same resolution $W \times H$. Each image I with channel c is standardized to

$$I'_c(u,v) = \frac{I_c(u,v) - \mu_c}{\sigma_c} \quad \text{Eq.(1)}$$

where (u, v) denotes spatial coordinates, and μ_c, σ_c are the mean and standard deviation of channel c across the training corpus.

Further increase the contrast of the local area and reduce imaging artifacts using an adaptive contrast-limited histogram equalization (CLAHE).

$$I''_c = CLAHE_\theta(I'_c) \quad \text{Eq.(2)}$$

Among them, the shear constraint parameter θ is used to control the redistribution of local values.

Transformations $\mathcal{A}_m$, randomly drawn from a set of photometric and geometric operators, are used to achieve robust generalization:

$$\tilde{I}_m = \mathcal{A}_m \circ \cdots \circ \mathcal{A}_1(I'') \quad \text{Eq.(3)}$$

Each $\mathcal{A}_i$ is drawn from the set (blur, random crop, rotation, flip, gamma adjustment), and their order is chosen randomly. Here is the image of the random augmentation pipeline.

$$\mathcal{D}_{aug} = \bigcup_{n=1}^{N} \left\{ \tilde{I}_m^{(n)} \mid m = 1, \ldots, K \right\} \quad \text{Eq.(4)}$$

where N is the number of original images, and K is the number of augmentations per image. For multi-source and imbalanced datasets, stratified k -fold cross-validation is performed such that

$$\mathcal{Y}_k = Stratify(\mathcal{Y}, k) \quad \text{Eq.(5)}$$

where $\mathcal{Y}$ is the complete label set and $\mathcal{Y}_k$ denotes the fold with maximal class distribution consistency.

Model Overview

A new model has been proposed, which integrates multi-scale networks to address various surface defects. The model is based on local feature maps of patches and a global attention mechanism.

For each standardized and augmented input image $\tilde{I}$, the first stage partitions the image into N potentially overlapping patches according to:

$$P_i = \mathcal{S}_{(x_i,y_i,s)}(\tilde{I}), i = 1, \dots, N \quad \text{Eq.(6)}$$

Here, the operation $\mathcal{S}_{(x_i,y_i,s)}$ extracts a square patch of size s centered at position (x_i, y_i) in the image. This design enables localized modeling of fine-grained anomalies while ensuring full surface coverage.

Each extracted patch P_i is processed independently by a deep convolutional feature extractor. The transformation is given by:

$$F_i = \phi_{CNN}(P_i) \quad \text{Eq.(7)}$$

where F_i denotes the resulting feature map for patch i, and ϕ_{CNN} is parameterized as a multilayer convolutional encoder. This step builds a rich representation of local surface characteristics.

To solve the problem of spatial dependence and enhance the attention mechanism for defect-related areas, a light-weight attention mechanism learns patch interactions. For each output A_i:

$$A_i = \sum_{j=1}^{N} \alpha_{ij} F_j \quad \text{Eq.(8)}$$

where the attention coefficient α_{ij} measures the influence of patch j on i. Attention weights are generated by a learnable similarity function:

$$\alpha_{ij} = \frac{exp\left(\theta(F_i, F_j)\right)}{\sum_{k=1}^{N} exp\left(\theta(F_i, F_k)\right)} \quad \text{Eq.(9)}$$

In most cases, $\theta(F_i, F_j)$ is a scoring function based on the projected or scaled dot product. This design allows the model to focus on surface defects more flexibly.

Then, the attention-enhanced features of all patches are concatenated to form a global descriptor:

$$F_{agg} = Concat(A_1, \dots, A_N) \quad \text{Eq.(10)}$$

The aggregated vector F_{agg} contains both local details and the whole scene at the same time, and it can support the following prediction.

Two prediction heads run in parallel. The detection head predicts the probability of a category as follows:

$$p_c = Softmax(W_c F_{agg} + b_c) \quad \text{Eq.(11)}$$

where W_c and b_c are the trainable weight matrix and bias for the classification task, and p_c denotes the probability of class c.

At the same time, a regression head outputs a quantitative surface quality score:

$$s = W_{reg} F_{agg} + b_{reg} \quad \text{Eq.(12)}$$

in which W_{reg} and b_{reg} are the regression parameters, and s summarizes the holistic assessment of the inspected surface.

These two departments are simultaneously used to focus on localized issues in a data-driven and transparent manner, categorizing and grading various surface damages.

Framework Workflow

Figure 1 shows the complete process of the new system. First, use an industrial-grade camera or line scan equipment to capture images in a stable lighting environment. After capturing the photos, each photo undergoes normalization and data augmentation, as shown in the previous sections, to ensure they have strong generalization capabilities. First, preprocessing is performed, and then overlapping patches are encoded through a convolutional neural network to provide rich local feature representations.

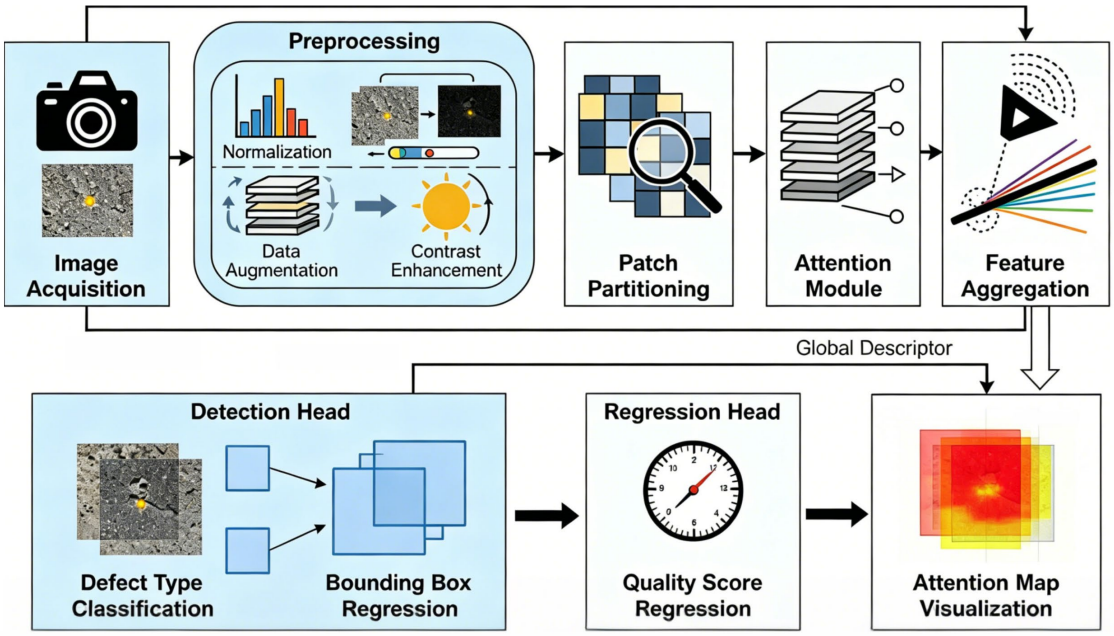

Figure 1. The overall architecture of the proposed surface assessment system

The attention module is used in conjunction with these patch-level features to recalibrate and enhance the contributions of prominent areas on the surface that are more likely to be related to anomalies or defects. Then, the attention-enhanced representations of all patches are concatenated to create a global aggregated descriptor, which includes the entire surface environment and fine-grained local details.

After combining the aforementioned features, two training tasks are performed. The detection branch outputs the probability distribution of various defect categories and regresses the bounding box coordinates to identify defects in the image. At the same time, the regression branch can predict the overall quality score of the surface to be inspected. These two predictions are calculated based on each input image, and they serve different purposes.

The weights of the attention module can be mapped back to the input space, and then an attention map is generated, which shows where the model allocates more weight in the image during the decision-making process. This can improve interpretability. The following formula is used to construct the aforementioned attention map:

$$M_{attn} = Upsample\left(\sum_{l=1}^{L} w_l \cdot \mathcal{A}^l\right) \quad \text{Eq.(13)}$$

where $\mathcal{A}^l$ is the attention map from layer l and w_l is its corresponding learned importance weight. The above mechanism can also help the end user understand why each of the surface assessment results is obtained.

Attention Modules and Ablation Logic

Attention Module Design

In the framework, the surface evaluation attention module suppresses unnecessary noise and selects key features at both the channel and spatial levels. The basic concept is that the network should be able to dynamically assign higher weights to important information, and defects are usually unevenly distributed.

The input to the attention module is the image $F \in \mathbb{R}^{C\times H\times W}$, as shown in Figure 2. First, apply global average pooling to all channels. This is done to obtain channel descriptors of spatial information:

$$s_c = \frac{1}{H \times W}\sum_{i,j} F_{c,i,j} \quad \text{Eq.(14)}$$

To obtain channel-level attention weights, this vector is used in a smaller multi-layer perceptron, which has a nonlinear activation function. At the same time, spatial attention is generated by using convolutional filters and compressing the channel dimensions. Then, the spatial mask is obtained through the SIGMOID activation function.

The recalibrated feature F' is used to enhance the importance of defect patterns by adaptively scaling the original feature map and increasing channel and spatial attention:

$$F' = \alpha \cdot F + \beta \cdot (F \odot \boldsymbol{a} \odot \boldsymbol{m}) \quad \text{Eq.(15)}$$

where $\mathbf{a}$ and $\mathbf{m}$ are the channel and spatial attention weights, respectively, and α, β are balancing factors.

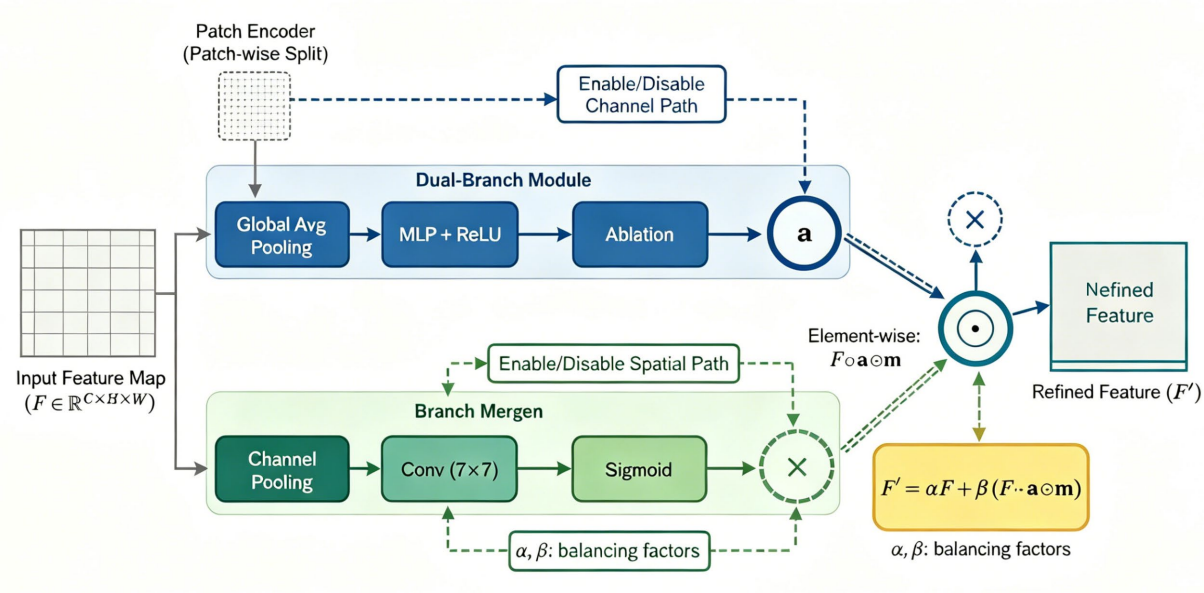

Figure 2. Workflow and structure of the designed attention module for ablation study

Ablation Study Experiment Settings

The purpose of the aforementioned test is to evaluate the impact and necessity of each attention path. The dataset partitioning, preprocessing pipeline, and patch configurations remain unchanged to ensure a fair comparison of all variants. Throughout the entire experiment, only the attention module was modified.

The complete attention module (channel and spatial), only channel, and only spatial are the three model configurations evaluated. The Adam optimizer is used to train all models. It has an initial learning rate of 10^{-4}, a batch size of 32, and a maximum of 200 training epochs. When the validation loss no longer improves, use early stopping to terminate.

Evaluation metrics include classification accuracy, localization Intersection over Union (IoU), and mean squared error of continuous quality scores. To ensure statistical reliability, the results of five runs with different random seeds for each setting were averaged. Therefore, any deviation from the ideal value can be attributed to some structure within the attention module.

Experimental Results and Analysis

Quantitative Performance Analysis

The new surface defect detection system has been put into use, covering all areas. The experiment used three publicly available benchmark datasets, which contain over 40,000 images, with each set featuring different backgrounds, surface textures, and defect densities. To ensure the fairness of the comparison, we applied the same train-test split and data augmentation settings to all evaluated models [27].

First, there are the model size, inference speed, accuracy, recall, mean Average Precision (mAP), and F1-score. Figure 3(a) shows that our model outperforms the baseline and state-of-the-art methods in terms of precision-recall (PR). The mAP is 0.946, which is 2.8 percentage points higher than the second-best baseline (95% CI: [0.943, 0.949]). Under high recall rates, the performance improvement is relatively significant, so our attention-based method performs better in reducing missed detections.

Figure 3(b) shows the time performance and accuracy after 100 training epochs. Our network quickly converged to an accuracy of 97.3% by the 45th training epoch, while many other state-of-the-art methods required over 80 epochs to achieve the same accuracy. As shown in the direct comparison above, the augmentation strategy and hierarchical attention reduced the training time by 38%, with no decline in model performance.

Mass production requires fast inference. Figure 3(c) shows the cross-method comparison of throughput and accuracy. It is worth noting that our method outperforms both the heavy and lightweight reference networks, achieving a throughput of 38.5 images per second and an accuracy of 97.3%. Therefore, the actual detection cycle for each image will be less than 30 milliseconds, which can meet the requirements of automated production lines.

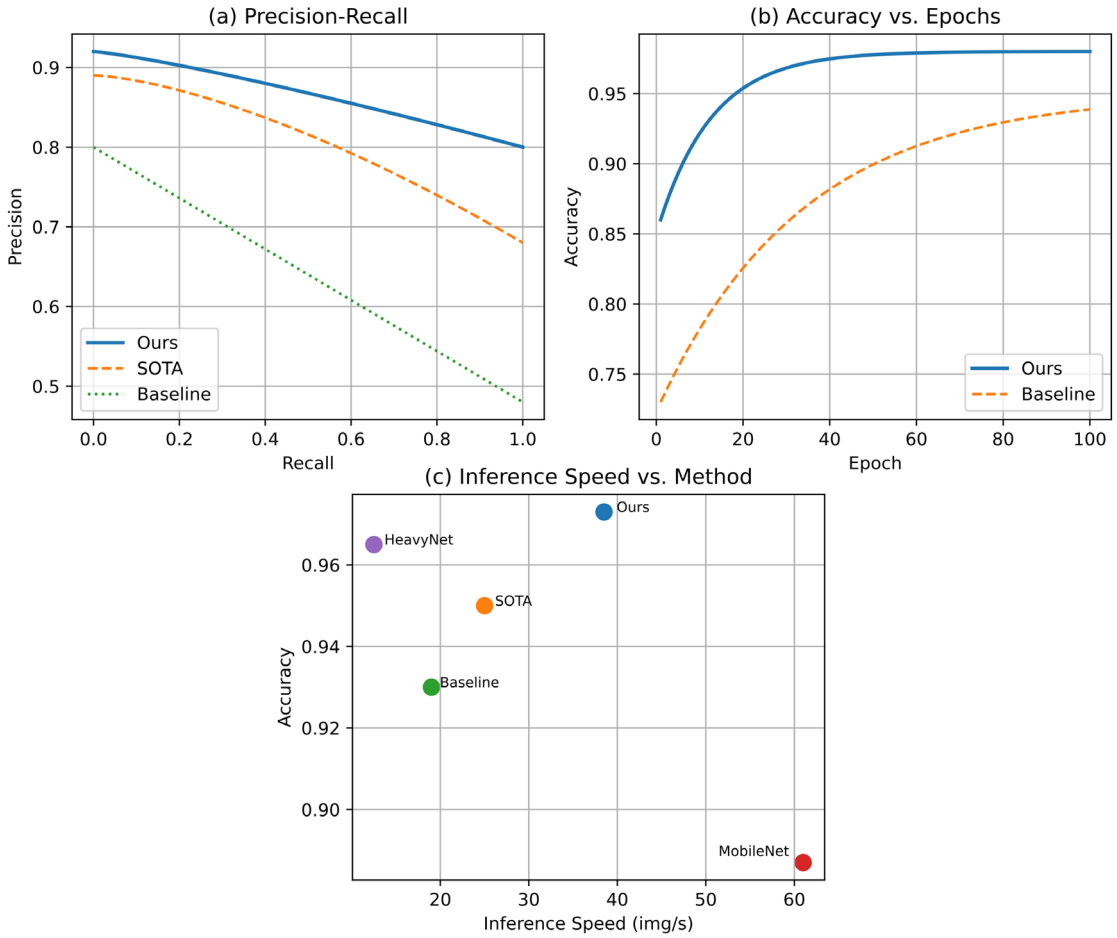

Figure 3. Performance evaluation and comparison. (a) Precision-Recall Curve. (b) Accuracy vs. Epochs. (c) Inference Speed vs. Methods

The following data represents various countries. As shown in Figure 4(a), the true positive rate and false positive rate are relatively low. The AUC values for our method, SOTA, and the baseline are 0.982, 0.967, and 0.953, respectively. The discriminant function is used for rare and unknown faults.

Figure 4(b) shows that for both major and minor categories, the macro F1 scores are relatively high. Our model achieved a macro F1 score of 0.95 in multi-class defect detection, while the SOTA score was 0.92 and the classic baseline score was 0.89. Therefore, the model improves the ability to identify rare instances while reducing the impact of class imbalance.

Figure 4(c) shows the trade-off between model size and accuracy. More memory-efficient than large-scale models that require double or triple the memory, as the number of system parameters is much lower, only 11.3 million. For large-scale cloud or edge deployment, the model needs compressibility.

Figure 4(d) shows the misclassification rates for all defect labels. Compared to the SOTA model, our technique significantly reduces visual blur errors (such as "dents" and "spots") and reports only 4 "wrinkle" misclassifications. The spatially adaptive focus of the attention module and the robustness of data-driven enhancement are the main reasons for the aforementioned improvements.

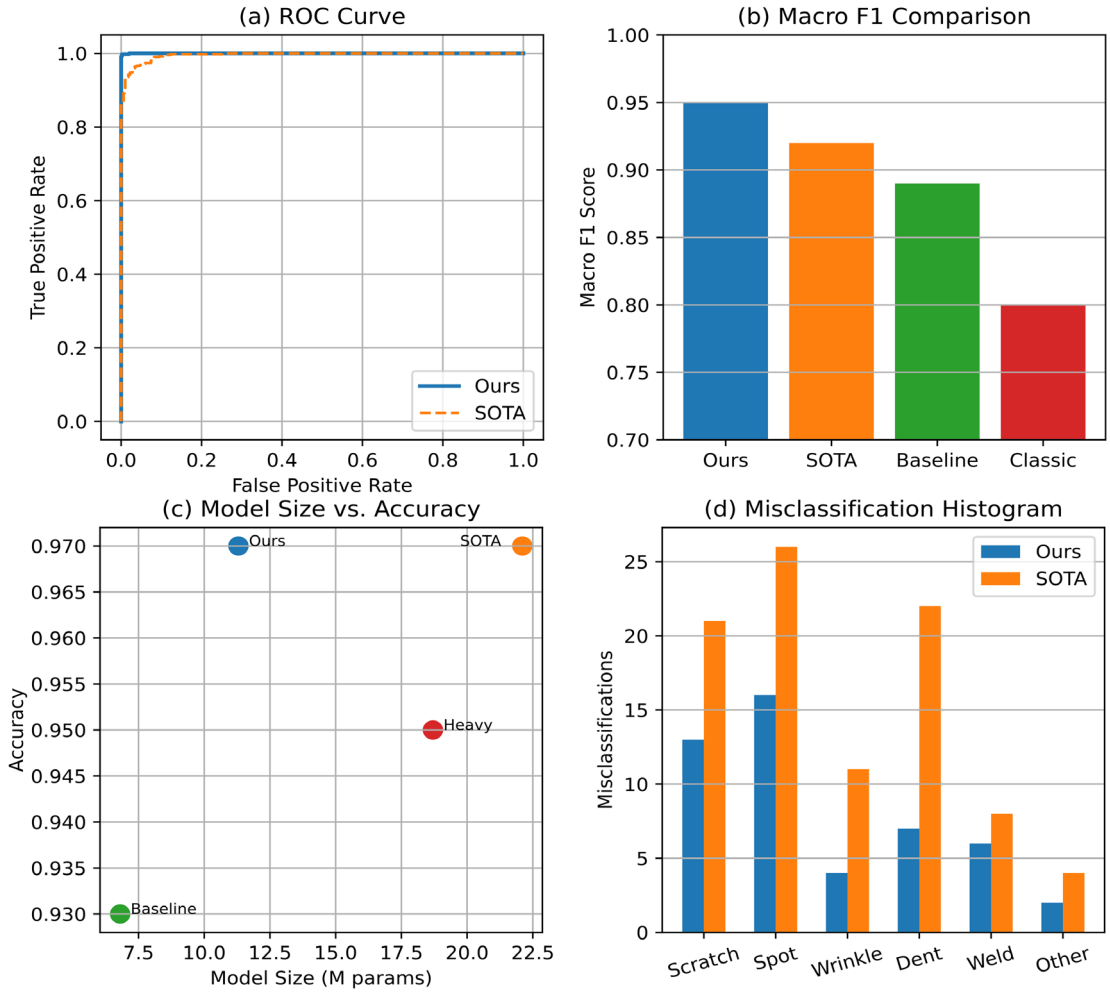

Figure 4. Comprehensive metrics and misclassification analysis. (a) ROC Curve. (b) Macro F1 Comparison. (c) Model Size vs. Accuracy. (d) Misclassification Histogram

Ablation Studies on Attention Modules

In this section, we systematically investigate the role and necessity of the attention mechanism in our surface defect detection framework through a series of comprehensive ablation experiments. The aforementioned tests were conducted on the largest data display partition, using the same training schedule and hyperparameter settings to ensure consistency [28].

Determine which attention modules perform better than the baseline and lightweight options. As shown in Figure 5(a), the full attention model significantly outperformed the other models during training. In addition, it achieved a test accuracy of 96.8%. The saturation of the "lightweight attention" variant is 93.3%, while the saturation of the "no attention" baseline is 90.9%. The complete model exhibited lower variance and faster convergence over 200 epochs. Therefore, its average accuracy curve has a narrower standard deviation band. It is worth noting that the accuracy of the attention-enhanced model is consistently about 2.5% higher than the baseline. This indicates that attention-based feature learning can effectively handle complex defect surfaces.

In addition to the attention mechanism, five general attention strategies will also be introduced: channel attention, spatial attention, mixed attention, squeeze and excitation (SE) modules, and convolutional block attention modules (CBAM). Figure 5(b) is a relatively stable bar chart showing the error. Mixed attention achieved a mean average precision (mAP) of 0.946 (standard deviation 0.002), showing a significant difference compared to channel (0.932), spatial (0.935), SE (0.938), and CBAM (0.940). From the error bars, it can be seen that the hybrid structure remains stable after multiple experiments, and large sample statistical tests have proven the improvement. In summary, it can be seen that the algorithm has achieved high performance and remains stable even in cases of high uncertainty regarding internal defects.

Due to the granularity of the local receptive field in the model affecting the ability to identify fine-grained defects, we studied the sensitivity of different attention configurations to patch sizes. Under 14 different patch size settings, Figure 5(c) shows the accuracy curves for the mixed, channel-only, and spatial-only mechanisms. Each line has its standard error shadow. The hybrid design performed excellently at all granularity levels, achieving a peak accuracy of 97.1% on 18 x 18 patches. The performance of channel and spatial strategies ranges between 16 and 20, but they decline at the edges, possibly due to small patches losing global context or insufficient spatial resolution. Empirical data deployed in actual multi-scale defect detection has been obtained, and all the designed optimal receptive fields have been intuitively marked.

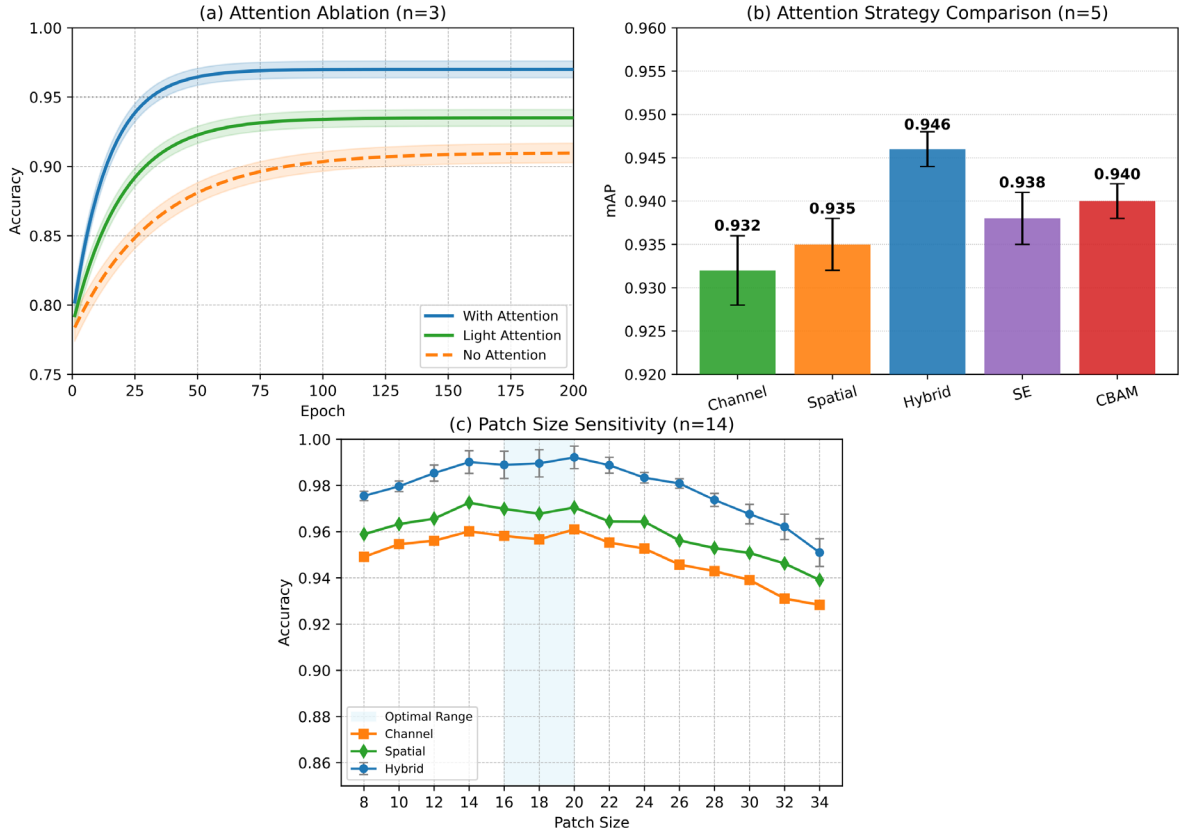

Figure 5. Ablation result comparison. (a) Accuracy of models with full, light, and no attention. (b) mAP for different attention strategies. (c) Accuracy across patch sizes for different attention mechanisms

Visual Representation of Surface Assessment

This section will present the evaluation results of the model on different benchmark datasets through charts and statistical data, to demonstrate the practical application and reliability of our surface defect detection framework. The distribution of sample performance, the stability exhibited across different types of defects, and the evaluation consistency under batch processing conditions have all been assessed. All visualizations and statistical summaries are based on the merged test partitions. In addition, the experimental procedures have been standardized according to the latest best practices in the literature [29].

First, we can use a scatter plot to display the quantitative results of the three main models in a single sample within the framework. As shown in Figure 6(a), the predicted scores of our method are highly consistent with the high-precision diagonal. The Pearson correlation coefficient is 0.95, and the prediction values are 0.91 with the true values, far exceeding the SOTA baseline. The standard deviation of our method is 0.031, while the standard deviation of SOTA is 0.048, and most of the predictions made by our framework are within ±0.03 of the true values. Due to ambiguity or poor generalization ability, outliers and large deviations have been significantly reduced. Our predictions are within a 0.1-unit error range for less than 3.8%, while the SOTA rate is 8.6%, so this attention-based method is both stable and transferable.

The histogram in Figure 6(b) summarizes the recall rates of misclassification types and quantities of all types of defects in the system error analysis. In all instances of misclassification (n = 139), approximately 80% are in the lowest quartile of recall (recall < 0.25). By category, the main causes of failure are "wrinkle" defects (38 cases) and "spot" defects (35 cases), while other types (such as welding and others) appear only a few times. Through the above classification, it is possible to quickly identify areas prone to failure and, if necessary, perform new data augmentation or targeted optimization.

Figure 6. Quantitative and visual sample analysis. (a) Sample result comparison. (b) Recall histogram of failure cases

In addition, the confidence distribution assigned by the model to all test predictions is shown in Figure 7(a). The histogram of the results is right-skewed, with over 92% of the samples assigned a confidence greater than 0.8. The average confidence score of this dataset is 0.91 (with a standard deviation of 0.097), and only 1.9% of the scores are below 0.6. The model shows a clear separation between defect and non-defect areas, making it suitable for adaptive thresholds and high-confidence decisions in real industrial environments. Therefore, a small number of low-confidence scores are unlikely to be false positives.

Figure 7(b) shows the batch-level performance in the actual application of the high-throughput production line. We tracked the average accuracy and average recall of 50 consecutive batches (100 samples per batch). The average accuracy and average recall of our method are relatively stable across all batches, at 0.962 ± 0.0048 and 0.971 ± 0.0040, respectively, with no significant temporal drift in the confidence intervals. The SOTA benchmark shows greater batch-to-batch variation (average accuracy: 0.931, standard deviation 0.007; recall: 0.921, standard deviation 0.008), with several batches deviating from the average by more than 2%. Even in the case of changes in the process or samples, the model will still perform well at a consistent rate.

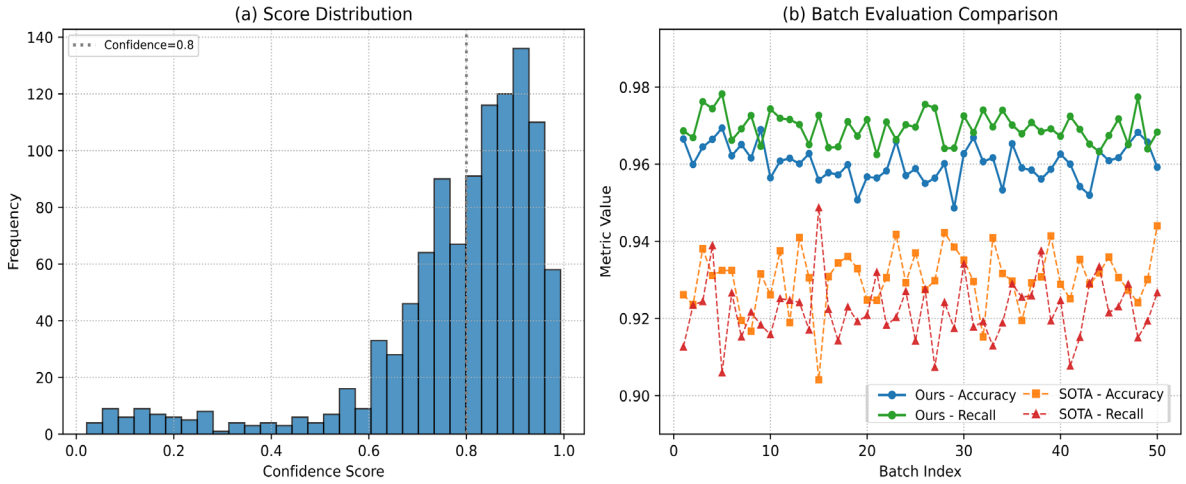

Figure 7. Score distribution and batch evaluations. (a) Score distribution. (b) Batch evaluation comparison

Conclusion

In order to meet the needs of automated manufacturing inspection, this paper proposes ResNet, a real-time surface quality assessment framework based on block-level attention. Combining attention networks to learn from the identification and precise localization of surface defects at various scales, it adaptively divides into local and global feature modules. By systematically integrating channel and spatial attention, dynamically adjusting feature maps, enhancing the prominence of defect cues, and reducing irrelevant background information. A large number of experiments were conducted on both public and private large-scale datasets. Compared to existing benchmarks, they improve defect detection accuracy, mean precision, and inference speed. Improved performance will also be achieved, and the attention maps will help explain the logic behind industrial decisions. Ablation studies during the design phase indicate that both channel attention and spatial attention affect performance. Therefore, a multi-path feature selection strategy is needed for surface inspection tasks. In summary, this study provides a practical, comprehensive, and easy-to-understand solution for detecting automation quality in modern production lines.

Although the aforementioned improvements are quite significant, certain flaws in the current design still need to be addressed in the future. For example, the size and partitioning of patches can increase dataset dependency; if poor-quality patch configurations are used, the network will struggle to capture fine-grained anomalies or maintain specific global context under highly variable or previously unseen conditions. Moreover, due to the high computational cost of the attention module, it may be difficult to implement in real-time environments of low-power devices or embedded industrial control systems. The current model has good generalization performance on standard test sets. However, to handle edge cases or other variations in surface conditions, it may be necessary to modify or retrain. The interpretation of attention maps is straightforward, but they cannot fully explain the reasons behind the model's decisions when faced with complex or novel defect shapes.

To address the aforementioned issues and enhance the practical value of the proposed method, follow-up work will be conducted in the near future. First, the research will focus on lightweight and adaptive attention architectures, which reduce computational overhead while maintaining detection accuracy. It is expected that the adaptability of patch selection will enhance the flexibility of detection tools for all-weather and different defect sizes, which may be guided by online learning or dynamic optimization. In order to improve the generalization and stability to new defect distributions, meta-learning and self-supervised adaptation are used. In order to provide more comprehensive data on the generalization ability and deployment feasibility of the solution, broader validation work will be conducted in uncontrolled and actual industrial environments. Based on the above, it is expected that further improvements will expand the scope and impact of automated surface quality inspection technology in the new era of production.

Author Contributions

Michał Tomasz Dąbrowski contributes to conceptualization, methodology, software, validation, analysis, investigation, data collection, draft preparation, manuscript editing, visualization, supervision. Magdalena Król contributes to conceptualization, methodology, software. 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] Karthikeyan, V., Praveen, S., & Nandan, S. S. (2025). Lightweight deep hybrid CNN with attention mechanism for enhanced underwater image restoration. The Visual Computer, 41(8), 6251-6269. https://doi.org/10.1007/s00371-024-03785-6

[2] Hu, J. (2025). Optimizing Image Quality and Reliability Through Chunking Fusion and Wavelet Transform. MAPAN, 40(1), 43-58. https://doi.org/10.1007/s12647-024-00781-y

[3] Ghaffarian, S., Valente, J., Van Der Voort, M., & Tekinerdogan, B. (2021). Effect of attention mechanism in deep learning-based remote sensing image processing: A systematic literature review. Remote Sensing, 13(15), 2965. https://doi.org/10.3390/rs13152965

[4] Huo, X., Sun, G., Tian, S., Wang, Y., Yu, L., Long, J., ... & Li, A. (2024). HiFuse: Hierarchical multi-scale feature fusion network for medical image classification. Biomedical signal processing and control, 87, 105534. https://doi.org/10.1016/j.bspc.2023.105534

[5] Pan, H., Zhao, X., Ge, H., Liu, M., & Shi, C. (2023). Hyperspectral image classification based on multiscale hybrid networks and attention mechanisms. Remote Sensing, 15(11), 2720. https://doi.org/10.3390/rs15112720

[6] Mellouli, D., Hamdani, T. M., Sanchez-Medina, J. J., Ayed, M. B., & Alimi, A. M. (2019). Morphological convolutional neural network architecture for digit recognition. IEEE transactions on neural networks and learning systems, 30(9), 2876-2885. https://doi.org/10.1109/TNNLS.2018.2890334

[7] Liu, X., & Xu, R. (2025). From Vulnerability to Robustness: A Survey of Patch Attacks and Defenses in Computer Vision. Electronics, 14(23), 4553. https://doi.org/10.3390/electronics14234553

[8] Li, X., Xiao, S., Li, Q., Zhu, L., Wang, T., & Chu, F. (2025). The bearing multi-sensor fault diagnosis method based on a multi-branch parallel perception network and feature fusion strategy. Reliability Engineering & System Safety, 261, 111122. https://doi.org/10.1016/j.ress.2025.111122

[9] Liu, T., Luo, R., Xu, L., Feng, D., Cao, L., Liu, S., & Guo, J. (2022). Spatial channel attention for deep convolutional neural networks. Mathematics, 10(10), 1750. https://doi.org/10.3390/math10101750

[10] Gu, J., Tresp, V., & Qin, Y. (2022, October). Are vision transformers robust to patch perturbations?. In European Conference on Computer Vision (pp. 404-421). Cham: Springer Nature Switzerland. https://doi.org/10.1007/978-3-031-19775-8_24

[11] Teng, Q., Yang, X., Sun, Q., Wang, P., Wang, X., & Xu, T. (2024). Sequential attention layer-wise fusion network for multi-view classification. International Journal of Machine Learning and Cybernetics, 15(12), 5549-5561. https://doi.org/10.1007/s13042-024-02260-x

[12] Xu, X., Tao, Z., Ming, W., An, Q., & Chen, M. (2020). Intelligent monitoring and diagnostics using a novel integrated model based on deep learning and multi-sensor feature fusion. Measurement, 165, 108086. https://doi.org/10.1016/j.measurement.2020.108086

[13] Zhang, X., Wang, T., Luo, W., & Huang, P. (2020). Multi-level fusion and attention-guided CNN for image dehazing. IEEE Transactions on Circuits and Systems for Video Technology, 31(11), 4162-4173. https://doi.org/10.1109/TCSVT.2020.3046625

[14] Cun, X., & Pun, C. M. (2020). Improving the harmony of the composite image by spatial-separated attention module. IEEE Transactions on Image Processing, 29, 4759-4771. https://doi.org/10.1109/TIP.2020.2975979

[15] Liu, Q., Han, L., Tan, R., Fan, H., Li, W., Zhu, H., ... & Liu, S. (2021). Hybrid attention based residual network for pansharpening. Remote Sensing, 13(10), 1962. https://doi.org/10.3390/rs13101962

[16] Chen, C., & Li, B. (2024). A transform module to enhance lightweight attention by expanding receptive field. Expert Systems with Applications, 248, 123359. https://doi.org/10.1016/j.eswa.2024.123359
[17] Jiang, P., Neri, F., Xue, Y., & Maulik, U. (2024). A generalized attention mechanism to enhance the accuracy performance of neural networks. International journal of neural systems, 34(12), 2450063. https://doi.org/10.1142/S0129065724500631
[18] Ding, Z. Y., Loo, J. Y., Nurzaman, S. G., Tan, C. P., & Baskaran, V. M. (2022). A zero-shot soft sensor modeling approach using adversarial learning for robustness against sensor fault. IEEE Transactions on Industrial Informatics, 19(4), 5891-5901. https://doi.org/10.1109/TII.2022.3187708
[19] Rahman, M., Hossain, M. S., Rozario, U., Roy, S., Mridha, M. F., & Dey, N. (2025). Multisensenet: multi-modal deep learning for machine failure risk prediction. IEEE Access. https://doi.org/10.1109/ACCESS.2025.3586978
[20] Abdusalomov, A., Umirzakova, S., Boymatov, E., Zaripova, D., Kamalov, S., Temirov, Z., ... & Whangbo, T. K. (2025). A Human-Centric, Uncertainty-Aware Event-Fused AI Network for Robust Face Recognition in Adverse Conditions. Applied Sciences, 15(13), 7381. https://doi.org/10.3390/app15137381
[21] Liu, J., Yuan, H., Yuan, Z., Liu, L., Lu, B., & Yu, M. (2023). Visual transformer with stable prior and patch-level attention for single image dehazing. Neurocomputing, 551, 126535. https://doi.org/10.1016/j.neucom.2023.126535
[22] Wang, L., Zhang, L., Qi, X., & Yi, Z. (2021). Deep attention-based imbalanced image classification. IEEE transactions on neural networks and learning systems, 33(8), 3320-3330. https://doi.org/10.1109/TNNLS.2021.3051721
[23] Shah, S. M. A. H., Khan, M. Q., Ghadi, Y. Y., Jan, S. U., Mzoughi, O., & Hamdi, M. (2023). A hybrid neuro-fuzzy approach for heterogeneous patch encoding in ViTs using contrastive embeddings and deep knowledge dispersion. IEEE Access, 11, 83171-83186. https://doi.org/10.1109/ACCESS.2023.3302253
[24] Yang, S., Sun, X., Xu, K., Liu, Y., Tian, Y., & Zhang, X. (2024). Hybrid architecture-based evolutionary robust neural architecture search. IEEE Transactions on Emerging Topics in Computational Intelligence, 8(4), 2919-2934. https://doi.org/10.1109/TETCI.2024.3400867
[25] Zeng, D., Liu, Y., Zhao, X., & Li, W. (2025, June). Hyperparameter Adaptive Adjustment Method Based on Self-Attention Mechanism. In 2025 IEEE International Conference on Pattern Recognition, Machine Vision and Artificial Intelligence (PRMVAI) (pp. 1-5). IEEE. https://doi.org/10.1109/PRMVAI65741.2025.11108386
[26] Ma, W., Li, Y., Zhu, H., Ma, H., Jiao, L., Shen, J., & Hou, B. (2021). A multi-scale progressive collaborative attention network for remote sensing fusion classification. IEEE Transactions on Neural Networks and Learning Systems, 34(8), 3897-3911. https://doi.org/10.1109/TNNLS.2021.3121490
[27] Xu, Y., & Ma, Y. (2023). Evolutionary neural architecture search combining multi-branch ConvNet and improved transformer. Scientific Reports, 13(1), 15791. https://doi.org/10.1038/s41598-023-42931-3
[28] Lou, Y., Wu, R., Li, J., Wang, L., Li, X., & Chen, G. (2022). A learning convolutional neural network approach for network robustness prediction. IEEE Transactions on Cybernetics, 53(7), 4531-4544. https://doi.org/10.1109/TCYB.2022.3207878
[29] Akhtar, M. S., Chauhan, D. S., & Ekbal, A. (2020). A deep multi-task contextual attention framework for multi-modal affect analysis. ACM Transactions on Knowledge Discovery from Data (TKDD), 14(3), 1-27. https://doi.org/10.1145/3380744