DEVELOPING AN AUTOMATED DAMAGE JUDGMENT SYSTEM FOR BRIDGE INSPECTIONS USING DEEP LEARNING

TAIKI FUJIMURA † 1, MASAKAZU MORIMOTO † 2 and TETSUNORI FUJITA ‡

† Graduate School of Engineering, University of Hyogo, Japan

‡ Hyogo Construction Technology Center for Regional Development

E-MAIL: † 1eg26r118@guh.u-hyogo.ac.jp, † 2morimoto@eng.u-hyogo.ac.jp

Abstract:

This study develops an automated system for judging the degree of bridge damage, alleviating the heavy burden that manual visual inspections place on administrators who process vast amounts of photographic data. We propose a two-stage deep learning system focusing on damage characteristics. Various methods are explored for comparative validation, including data cleansing using Generative AI and image preprocessing. The system improved overall accuracy to 0.90, notably enhancing recall for the critical damage category 'e' from 0.55 to 0.75.

Keywords:

Bridge Inspection; Deep Learning; Generative AI; Two-Stage Classification

1. Introduction

The aging of infrastructure, such as bridges constructed during periods of rapid economic growth, has become a severe social problem. In Japan, the rapid increase of bridges exceeding 50 years old—known as the "2033 problem"—makes appropriate maintenance and regular inspections essential to prevent collapses.

While mandated close visual inspections every five years ensure safety, the process of verifying vast amounts of photographic data during the scrutiny of inspection records—in addition to the inspections themselves—are highly labor-intensive amid a severe shortage of engineers. Therefore, drastic labor-saving in these administrative tasks is essential for improving operational efficiency.

While numerous deep learning methods have been proposed for automating concrete damage detection, most research focuses on detecting damage presence and location[1], [2], [3], [4]. In contrast, accurately classifying damage progression from noisy, real-world inspection images remains a significant challenge.

To address this issue, this study aims to develop an automated damage judgment system for bridge inspection images. Focusing on concrete spalling and rebar exposure, we propose a data cleansing method using multimodal Generative AI and a two-stage deep learning system combining background removal and feature extraction, subsequently verifying their classification accuracy.

The remainder of this paper is organized as follows: Section 2 details the proposed methodology, Section 3 presents the experimental results, and Section 4 concludes the study.

2. Proposed Methodology

2.1. Dataset Construction and Data Cleansing

In this study, an initial dataset was constructed by automatically extracting images from bridge inspection records. However, a major challenge was encountered: as shown in Figure 1, the extracted data contained not only appropriate damage images like (a), but also many inappropriate images, such as distant shots where the damaged areas are indistinguishable, like (b). These noisy images severely degrade the classification accuracy of deep learning models.

To address this issue, we propose a data cleansing method utilizing a multimodal Generative AI (Gemini 2.5 Flash)[5]. As shown in Figure 2, we provided the AI with not only the inspection images but also the corresponding text from the memo field. A customized prompt was engineered to evaluate the suitability of each image based on both visual and textual information. While a comparative analysis with other state-of-the-art models was not conducted due to resource constraints, it remains a key objective for future work. The primary instructions are as follows:

- **Keep:** Close-up images where the concrete surface and its condition are clearly distinguishable.

(a) Appropriate close-up image (b) Inappropriate distant image

FIGURE 1. Examples of extracted images

Photo No.	39
Span No.	1
Member Name	Main Girder
Member No.	16
Damage Type	⑦Spalling・Rebar Exposure
Condition State	c
Memo	

【Previous Inspection (2016) → Current Inspection】
⑦ Spalling・Rebar Exposure：c→a (Re-evaluated)
The evaluation category of the photo differs from the actual evaluation.

FIGURE 2. Examples of an inspection image and its corresponding memo

- **Discard:** Distant images where the concrete surface or damaged areas cannot be visually confirmed, or noisy images with notes such as "already repaired" in the memo field.

By having the AI interpret both visual features and textual information, this method was designed to effectively filter out inappropriate images.

2.2. Evaluation Categories for Spalling and Rebar Exposure

In this study, we focus on 'spalling and rebar exposure,' which are critical deterioration factors in concrete bridge components. Based on the Bridge Periodic Inspection Guidelines of the Ministry of Land, Infrastructure, Transport and Tourism (MLIT), this damage is classified into four evaluation categories (a, c, d, and e, omitting category b as per MLIT guidelines) according to the degree of progression. The evaluation criteria for each category are presented in Table 1, and examples of corresponding images are shown in Figure 3. The deep learning model developed in this study takes inspection images as input and aims to automatically classify them into these four classes. Therefore, each image in the dataset is assigned an evaluation category as a ground truth label, based on the records from the inspection reports created by experienced inspectors.

TABLE 1. Evaluation criteria for spalling and rebar exposure

Evaluation Category	Evaluation Criteria
a	No damage (Including repair marks)
c	Spalling only
d	Rebar exposure with minor corrosion
e	Rebar exposure with severe corrosion

(a) Example of category a (b) Example of category c

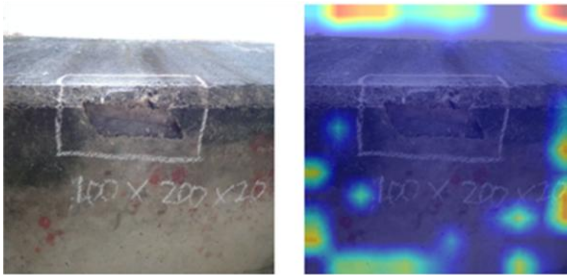
(c) Example of category d (d) Example of category e

FIGURE 3. Examples of images for each evaluation category

2.3. Single-Stage Classification and Background Removal

Initially, this study investigated a single-stage 4-class classification (a, c, d, e) method using the images that had undergone the data cleansing described in the previous section. Through preliminary experiments comparing several representative CNN(accuracy: 0.76) and ViT models, MaxViT-Base(accuracy: 0.84) was selected as the baseline model due to its superior capability in extracting both global and local features[6], [7]. However, when visualizing the reasoning behind the model's predictions using Grad-CAM, numerous instances were observed where the model incorrectly focused on background noise, leading to misclassifications (Figure 4)[8].

To correctly guide the model's attention to the actual damaged areas, we introduced background removal using U²-Net as a preprocessing step [9]. This model was trained using inspection records from other cities, independent of the main experiment. Mask images were created where chalked regions indicating damage were labeled white and the background black.



(a) Original Image (b) Grad-CAM Visualization

FIGURE 4. Visualization using Grad-CAM

(a) Original Image (b) Mask Image

(c) Cropped Image

FIGURE 5. Examples of background removal using U^2-Net

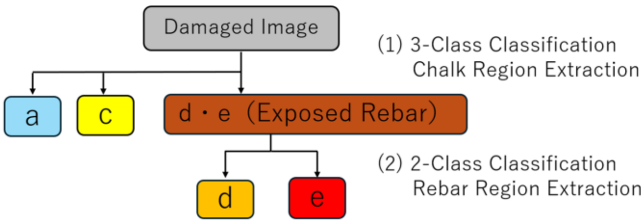


FIGURE 6. Processing flow of "two-stage classification method"

(a) Original Image (b) Cropped Image

FIGURE 7. Example of rebar region extraction

Using the masks automatically generated by the trained U^2-Net, the background was removed to crop out only the damaged areas. Furthermore, to standardize the image aspect ratio to 1:1, a padding process was applied by filling the removed background areas with the mean color of the image (Figure 5). This series of preprocessing steps was applied to the training dataset (evaluation categories c, d, and e), and in the next section, we conduct a comparative validation of the classification accuracy with and without this background removal.

2.4. Two-Stage Classification Method

Although the background removal method utilizing U^2-Net introduced in the previous section effectively guides the model's attention to the damaged areas, the accuracy of a single-stage 4-class classification saturated. As will be detailed in the experimental results (Section 3), a detailed analysis of the misclassifications revealed frequent confusion between evaluation categories "d" and "e," both of which involve exposed rebar.

To address this issue, this study proposes a "two-stage classification method" that focuses on the degree of corrosion progression as the primary difference between "d" and "e," classifying the targets through a stepwise refinement process. The specific processing flow of this method is illustrated in Figure 6.

In the first stage, a three-class classification is performed using the chalk-marked area images extracted by U^2-Net to distinguish among "a" (no damage), "c" (spalling only), and "d/e" (exposed rebar). In the second stage, for images classified as "d/e" in the first stage, U^2-Net is utilized again to extract only the rebar region (Figure 7), followed by a two-class classification based on the degree of corrosion. By optimizing the region of interest at each stage in this manner, it is expected that the classification accuracy between categories "d" and "e," which are traditionally difficult to distinguish, will be improved.

3. Experiments and Results

3.1. Evaluation of Data Cleansing with Generative AI

In this section, we evaluate the effectiveness of the data cleansing method using Gemini 2.5 Flash proposed in Section 2. The experiment utilized a total of 2,675 images automatically extracted from inspection records. A visual inspection was conducted to separate "close-up images" suitable for training data from inappropriate "distant images," resulting in 1,419 images being defined as the ground truth.

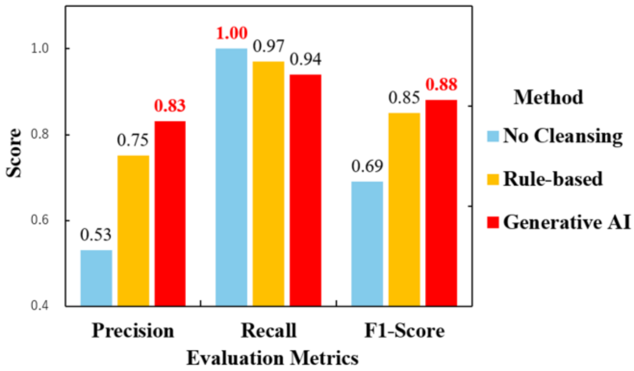


FIGURE 8. Comparison of cleansing methods

TABLE 2. Comparison of classification accuracy with and without background removal

Background removal	Accuracy	Macro-average F1-score
Without	0.84	0.79
With	0.86	0.82

For comparison, the extraction accuracy was evaluated under three conditions: "No Cleansing," a "Rule-based method" (which filters out distant images using only descriptions in the memo field), and the "Proposed Method." Precision, Recall, and F1-Score were used as evaluation metrics. The experimental results are shown in Figure 8.

As shown in Figure 8, the "No Cleansing" condition resulted in a precision of 0.53, indicating that approximately half of the dataset consisted of inappropriate images. The rule-based method, which relies solely on text keywords, improved precision to 0.75 but failed to completely eliminate noise due to variations in textual descriptions.

In contrast, the proposed method using Generative AI achieved the highest values with a precision of 0.83 and an F1-score of 0.88. Although the recall (0.94) was slightly lower compared to the rule-based method (0.97), this is considered to be the result of the AI strictly excluding ambiguous borderline images.

From these results, it can be concluded that while the proposed method does not achieve complete automation of dataset construction (i.e., a precision of 1.00), it is highly effective as an "auxiliary screening system." It significantly reduces noise from vast amounts of image data, thereby alleviating the burden of manual visual confirmation by humans.

Note that in the damage classification experiments in the next section, we utilize the high-quality dataset of 1,419 images, which has been finalized through visual confirmation, to accurately verify the performance of the models themselves.

3.2. Effectiveness of Background Removal and Limitations of Single-Stage Classification

In this section, we validate the effectiveness of the background removal using U²-Net proposed in Section 2 and clarify the challenges of single-stage 4-class classification using MaxViT-Base. For the initial experiment, we used a dataset of 1,419 images (a: 125, c: 493, d: 708, e: 93) constructed through the cleansing and visual confirmation processes described above. The model was evaluated using 5-fold cross-validation. To address the class imbalance problem common in damage assessment, data augmentation was applied to ensure the number of training images per class was equalized in each epoch[10].

First, to evaluate the effect of the preprocessing, we trained the 4-class classification models using both "original images without background removal" and "images with backgrounds removed by U²-Net, isolating only the damaged areas," and compared their classification accuracies. The results are shown in Table 2.

From Table 2, it can be confirmed that in the dataset where background removal was applied, the overall accuracy improved from 0.84 to 0.86, and the macro-average F1-score improved from 0.79 to 0.82. This improvement is considered to be because the model's focus was successfully guided away from classification-irrelevant background noise toward the actual damaged areas.

While background removal was effective, overall accuracy plateaued at 0.86. To improve performance, we refined the visual screening criteria and expanded the dataset to 1,921 images (a: 125, c: 1000, d: 650, e: 146). To maintain evaluation fairness, the additional images were strictly limited to an independent test set. These images were not used in the training of the classification models or U²-Net, thereby preventing data leakage and ensuring strict data independence.

As a result, although the accuracy slightly improved to 0.87, a plateau in performance was still observed. To identify the cause of this limitation, we generated a confusion matrix of the classification results and conducted a detailed analysis. The results are shown in Figure 9.

As is evident from the confusion matrix in Figure 9, although the classification accuracy for "a" (no damage) and "c" (spalling only) is high, misclassifications frequently occur between "d" (minor corrosion) and "e" (severe corrosion)—both of which involve exposed rebar—partly due to the small sample size of category "e." Specifically, numerous cases were observed where images of the severe damage category "e" were underestimated as "d," resulting in a recall of only 0.55 for "e.

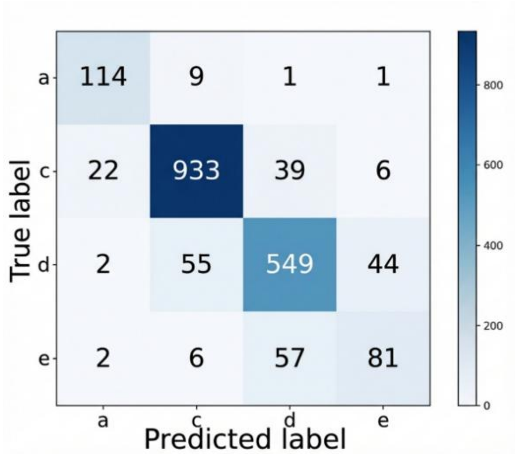


FIGURE 9. Confusion matrix of the 4-class classification

The textural features of the concrete surface differ significantly from the color space and morphological features of rust (corrosion) on the exposed rebar. Therefore, it is considered that there is a structural limitation in evaluating these distinct features simultaneously with high accuracy using a single network architecture.

To address this and improve detection of the critical category "e," the next section evaluates the proposed two-stage classification method.

3.3. Validation of the Two-Stage Classification Method

To address the issues identified in the previous section, this section evaluates the accuracy of the proposed "two-stage classification method." In the first stage, a 3-class classification (a: no damage, c: spalling, d/e: exposed rebar) was performed using MaxViT-Base on images with backgrounds removed via U²-Net. The evaluation employed 5-fold cross-validation, with data augmentation applied to balance the number of images across classes. The resulting confusion matrix is shown in Figure 10.

As a result of the experiment, both the overall accuracy and the macro-average F1-score reached 0.95, achieving extremely high classification accuracy. It can also be confirmed from Figure 10 that misclassifications are very few. This is considered to be because merging evaluation categories "d" and "e" into a single class eliminated the need for the model to identify fine features such as the degree of rebar corrosion, thereby reducing the difficulty of the classification task.

Based on these results, the model constructed in this experiment is adopted for the first stage (3-class classification) of the two-stage classification method.

Next, we conducted a comparative experiment to determine the optimal model construction method for the second stage of the "two-stage classification method," which involves the two-class classification of evaluation categories "d" and "e." Specifically, we evaluated the following three learning approaches: (1) Image-only: learning using MaxViT-Base with only the extracted rebar region images described in Section 2; (2) Feature-only: learning using Random Forest based solely on numerical features extracted from the images from four perspectives: color, texture, frequency, and shape; and (3) Hybrid: integrated learning combining both images and numerical features. In the training process, 5-fold cross-validation and data augmentation were applied.

In this stage, the most critical error to avoid is overlooking category "e," which represents severe damage. Therefore, rather than simply using accuracy, we employed recall—which indicates the infrequency of false negatives (overlooks)—and AUC (Area Under the Curve)—a comprehensive indicator of classification performance—as evaluation metrics. The comparison of each evaluation metric is shown in Figure 11.

As a result, the "hybrid approach" integrating images and features achieved an AUC of 0.90 and recorded a recall of 0.77 for the most important metric, evaluation category "e," demonstrating the highest discrimination performance among the compared methods. This improvement is attributed to the complementary effect of the abstract feature extraction by the CNN and the physical features represented by the numerical data, which contributed to enhancing the classification accuracy.

Based on these results, we concluded that the hybrid approach is the most suitable model for this task and adopted it for the second stage (two-class classification) of the two-stage classification method.

Finally, we compare the results of the 4-class classification using the two-stage classification method with those of the conventional single-stage 4-class classification. The confusion matrix obtained by the two-stage classification method is shown in Figure 12.

Comparing Figure 9 and Figure 12, the overall accuracy improved from 0.87 to 0.90. Furthermore, general improvements were observed in the precision, recall, and F1-score across each class. In particular, the recall for category "e," which had been a significant challenge, improved drastically from 0.55 to 0.75. This improvement is attributed to the division of the task into two distinct stages—the first stage focusing on the state of exposed rebar, and the second stage focusing on the state of corrosion. This division clarified the specific features the model needed to learn at each respective stage.

These results strongly demonstrate the effectiveness of the proposed two-stage classification method.

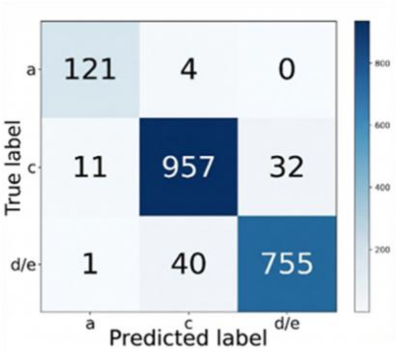


FIGURE 10. Confusion matrix of the first-stage 3-class classification

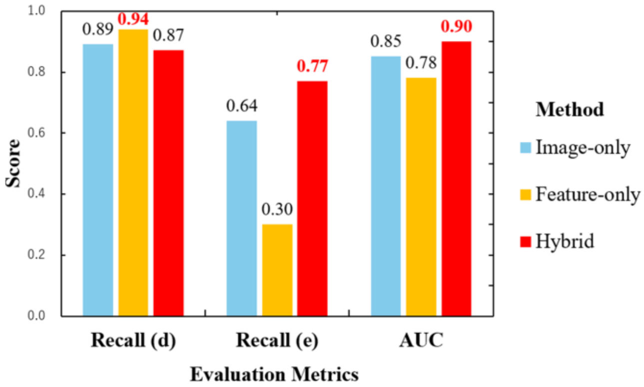


FIGURE 11. Comparison results of the three methods in the second stage (2-class classification)

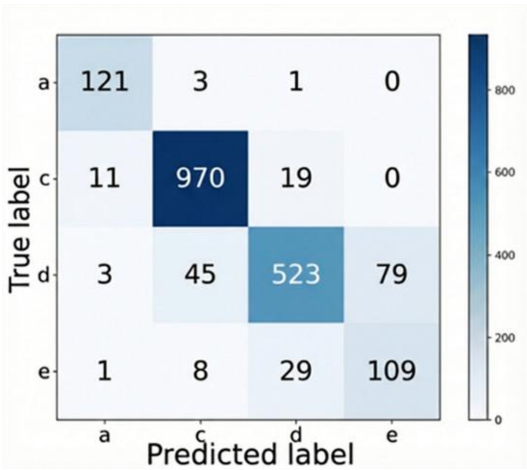


FIGURE 12. Confusion matrix of the two-stage classification method

4. Conclusion

In this study, to promote infrastructure DX, we developed an automated damage assessment system for "spalling and rebar exposure" using deep learning. First, by utilizing multimodal Generative AI for data cleansing, we effectively removed noise from the dataset and significantly reduced the labor required for manual visual confirmation.

For the damage classification task, since a single-stage four-class classification model struggled to distinguish between rebar exposure categories "d" and "e" (accuracy: 0.87), we proposed a "two-stage identification method". This method first identifies the presence of rebar exposure with an accuracy of 0.95 using background removal via U²-Net, followed by evaluating the degree of corrosion through a hybrid approach that integrates CNN-extracted features with physical numerical data.

As a result, the proposed method improved the overall classification accuracy from 0.87 to 0.90. Notably, the recall for Category "e"—the most critical level of severe damage—was drastically improved from 0.55 to 0.75. Future work includes further enhancing accuracy by expanding the training data for Category "e" and extending this automated assessment system to other major defects, such as cracks and water leakage, to contribute to the comprehensive efficiency of bridge inspection operations.

References

[1] B. F. Spencer Jr, et al., "Advances in computer vision-based civil infrastructure inspection and monitoring", *Engineering*, Vol 5, No. 2, pp. 199-222, 2019.

[2] M. Yoshikura, "Enhancement of visibility for damage detection of concrete bridges...", Ph.D. dissertation, Kanazawa Univ., 2023.

[3] Y. J. Cha, et al., "Deep Learning-Based Crack Damage Detection", *Comput. Aided Civ. Inf.*, Vol 32, No. 5, pp. 361-378, 2017.

[4] W. Xu, et al., "Bridge Crack Detection Based on Improved Faster R-CNN", *Applied Sciences*, Vol 9, No. 18, 2019.

[5] Gemini Team, Google, "Gemini 1.5: Unlocking multimodal understanding...", *arXiv:2403.05530*, 2024.

[6] K. He, et al., "Deep Residual Learning for Image Recognition", *Proc. CVPR*, pp. 770-778, 2016.

[7] Z. Tu, et al., "MaxViT: Multi-Axis Vision Transformer", *Proc. ECCV*, pp. 459-479, 2022.

[8] R. R. Selvaraju, et al., "Grad-CAM: Visual Explanations from Deep Networks", *Proc. ICCV*, pp. 618-626, 2017.

[9] X. Qin, et al., "U²-Net: Going deeper with nested U-structure", *Pattern Recognition*, Vol 106, p. 107404, 2020.

[10] T. Y. Lin, et al., "Focal Loss for Dense Object Detection", *Proc. ICCV*, pp. 2980-2988, 2017.