Extraction of Abnormal Rust Regions of Weathering Steel Using Semantic Segmentation

Shuta NOTSU [a], Makoto OHYA [b], Makoto FUJIU [c]

[a] *Graduate School of Geosciences and Civil Engineering, Kanazawa University, Kanazawa City, 920-1192, Japan; E-mail: p2307mctku@gmail.com*
[b] *Professor, Department of Civil and Environmental Engineering, National Institute of Technology Matsue College, Matsue City, 690-0865, Japan; E-mail: ohya@matsue-ct.ac.jp*
[c] *Professor, Faculty of Transdisciplinary Sciences for Innovation, Kanazawa University, Kanazawa City, 920-1192, Japan; E-mail: fujiu@se.kanazawa-u.ac.jp*
Same as the first author; E-mail: p2307mctku@gmail.com

Abstract: Weathering steel is a material with the unique characteristic that the rust layer formed on its surface sufficiently reduces the corrosion rate and controls rust with rust. Therefore, when inspecting weathering steel bridges, it is necessary to evaluate the condition of the rust. The basic method is to rate the grain size on a 5-point scale based on the appearance of the rust. Grades 3 to 5 are considered good and require no additional action, while grades 2 and 1 require observation and immediate action, respectively, and accurate judgment is required. However, since quantitative criteria are not always provided, even experts may sometimes find it difficult to make a definitive judgment. Therefore, the purpose of this study is to construct a system that uses deep learning to extract regions with abnormal rust on the weathering steel surfaces with non-uniform corrosion.

Keywords: Weathering Steel, Rust Appearance Evaluation, Semantic Segmentation

1. INTRODUCTION

In Japan, an increasing number of infrastructure structures have been constructed more than 50 years ago, and appropriate management to maintain their functions has become an important issue. Infrastructure structures must be visually inspected every five years, but there is a serious shortage of professional engineers to take on this task, and there is an urgent need to develop efficient maintenance management technology for the approximately 730,000 bridges that exist in Japan. In recent years, inspection images acquired by UAVs (unmanned aerial vehicles) and efficient inspection methods using image processing technology and machine learning have attracted attention, and the introduction of infrastructure inspection using AI technology is being actively promoted (Aoshima et al. 2018, Tabata et al. 2018, Goto et al. 2006).

Steel bridges account for about 38% of all bridges in Japan, or about 270,000 bridges. Steel bridges play an important role in Japan's infrastructure, and bridges using weathering steel, which has the feature of suppressing rust with rust, have attracted particular attention because of their potential to reduce life cycle costs. Weathering steel has the ability to sufficiently reduce the corrosion rate in an appropriate corrosive environment due to the rust layer that forms on the surface, making it possible to use this material without painting when used in steel structures. The percentage of weathering steel bridges in Japan is estimated to be about 11% according to the latest data.

Table 1. Rust appearance grading criteria

Grade	Rust thickness	Feature
5	Less than 200μm	Rust is minimal and the color is relatively bright.
4	Less than 400μm	Rust is fine and uniform with a size of less than 1 mm.
3	Less than 400μm	Rust is a course, ranging in size from 1-5 mm.
2	Less than 800μm	Rust is scaly, ranging in size from 5 to 25 mm.
1	More than 800μm	Rust has laminar exfoliation.

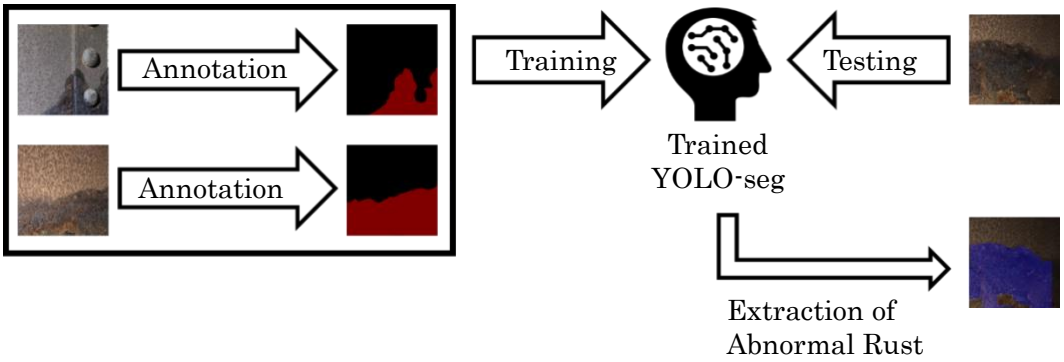


Figure 1. Training and validation process of the system

In the inspection of painted steel bridges, the corrosion protection performance is evaluated based on the degree of coating film deterioration, and the occurrence and extent of rust. On the other hand, in the case of weathering steel bridges, inspections are conducted to evaluate the state of protective rust that forms on the surface of steel materials, and appropriate repairs are made according to the state of the rust (Imai et al. 2012). A standard method for evaluating the condition of rust on weathering steel bridges is a five-point grading scale for grain size based on the appearance of the rust; a score of 3 to 5 is considered good and requires no additional action, while a score of 2 and 1 requires observation and immediate action, respectively. Table 1 shows the rust appearance grading criteria. However, these criteria are not necessarily quantitative, and since subjectivity tends to enter appearance evaluation, it is difficult for even experts to make accurate judgments. To address this issue, research is underway to achieve highly accurate identification of rust conditions in weathering steel bridges using AI technology (Tai et al. 2021). In inspections of weathering steel bridges, a system that can quantitatively extract areas of abnormal rust from the steel surface, where non-uniform rusting has occurred throughout, is desired.

In this paper, it is attempted to contract a system using semantic segmentation based on deep learning to evaluate quantitatively extracted areas of rust with abnormal rust from steel surfaces with non-uniform rust conditions. The objective of this study is to examine the difference in the accuracy of extracting abnormal rust due to the difference in the image size used in the dataset.

2. OVERVIEW OF THE ABNORMAL RUST REGION EXTRACTION SYSTEM

This study aims to construct a system for quantitatively extracting areas of abnormal rust from steel surfaces with non-uniform rust conditions. By utilizing semantic segmentation, one of the deep learning techniques, to classify rust conditions at the pixel level, a new inspection method is being developed that supports visual inspections. Figure 1 illustrates the training and validation process of this system, visualizing a series of steps including the preparation of

training data, model training, and evaluation of detection accuracy for each grade of rust. To build the system, a dataset was created by annotating areas of abnormal rust on images of weathering steel bridges with non-uniform corrosion. This dataset was used to train an AI model capable of performing semantic segmentation tasks. The AI model chosen was YOLO (You Only Look Once, Redmon et al. 2016) version 8's segmentation model (YOLOv8-seg). This model enables high-accuracy, real-time segmentation processing and is expected to improve the efficiency of infrastructure inspection when applied to inspection tasks. During training, cross-entropy loss was used as the loss function for YOLOv8-seg, allowing the model to learn to accurately classify the rust state of each pixel.

The developed system can effectively assist visual inspections by detecting the progress of rust in more detail and visualizing the state of deterioration at the pixel level. To verify the performance of the constructed system, evaluation data consisting of images of unevenly corroded weathering steel bridge surfaces not used in training were utilized as test images. The system's ability to appropriately extract areas of abnormal rust was examined.

3. BUILDING AN ABNORMAL RUST REGION EXTRACTION SYSTEM

3.1 Dataset Generation

This study focuses on unpainted weathering steel bridges with non-uniform corrosion. Five images of weathering steel bridges with non-uniformed corrosion, each 3024x4032x3 pixels, were used as original images. source images. Four were used for training, and one for system evaluation. From the original images, three types of smaller region images measuring 320x320x3, 640x640x3, and 1280x1280x3 pixels were extracted using shifting techniques. Each dimension represents height, width, and RGB respectively. To investigate the impact of image size on training, datasets were created by annotating areas of abnormal rust on images of three sizes: the standard 640x640 pixels used with YOLO-seg, as well as images with half and double the standard height and width. Since abnormal rust in the images of the weathering steel bridge used in this study is caused by leakage, the annotated labels were grade 1 for areas of rust with grades 1 and 2, and background for areas of rust with grades 3 to 5. Images that showed no corrosion or did not contain both normal and abnormal rust were not used for training and validation. To expand the number of datasets, annotated images were prepared that were inverted vertically, horizontally, and both vertically and horizontally. Figure 2 shows a portion of the three types of small region images and annotation data used for training. Table 1 shows

(a) 320x320 (b) 640x640 (c) 1280x1280

Figure 2. Small area images used for training and some of the annotated data

Table 2. The number of data (training and validation sets) used to train YOLOv8-seg

	Train data	Validation data	Test data (small area)	Test data (original image)
320x320pixels	400	100	80	1
640x640pixels	185	47	26	1
1280x1280pixels	96	24	8	1

the number of data used for YOLO-seg training (training and validation sets) and the number of test data used to evaluate the abnormal rust detection accuracy of the trained YOLO-seg. The ratio of training to validation data was set at 4:1. For the test data, both the original images and the image sizes used for training were applied.

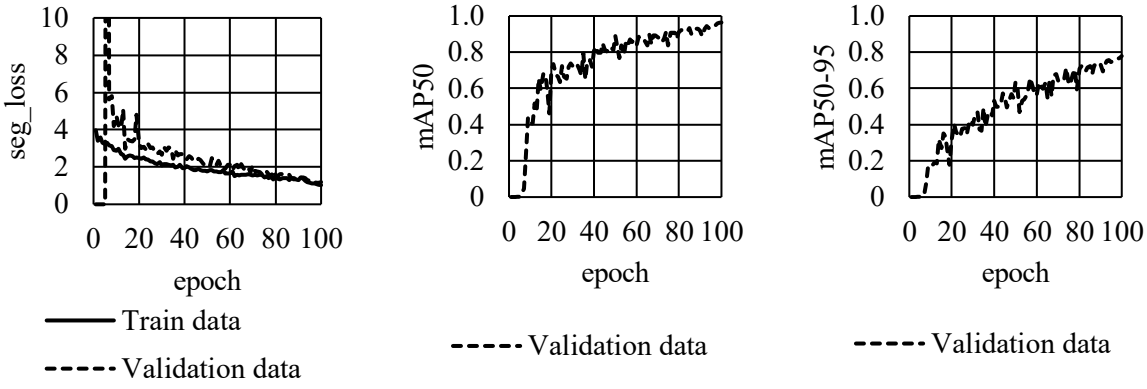
3.2 Train and Validation Result

In this study, Train data was used to build the learning model, while Validation data was used to evaluate the performance of the learned YOLOv8-seg model. To evaluate the learning progress, the number of learning sessions (epochs) was set to 100, and performance during the learning process was checked sequentially.

The learning progress was evaluated using two indices: seg_loss (Segmentation Loss), which indicates the detection ability in the segmentation task, and mAP (Mean Average Precision), which indicates the overall accuracy of the prediction mask. Training status of YOLOv8-seg by Train data and validation status of trained YOLOv8-seg by Validation data are summarized in Fig. 3 to Fig. 5. Figures 3 to 5 show the training results for 320x320, 640x640, and 1280x1280, respectively. Also shown are (a) seg_loss per epoch, (b) mAP50 per epoch in the Validation data, and (c) mAP50-95 per epoch in the Validation data. seg_loss is in the segmentation task, The loss function is the error between the predicted mask and the correct mask. The smaller this value is, the more accurate the model is at segmenting the input data. A decrease in seg_loss during the learning process is an important indicator of improved segmentation performance because it suggests that the model can accurately identify the contours and internal structure of the object. mAP50 is an index that represents the average prediction accuracy when the overlap between the correct and predicted regions is 50% or more. This is a relatively loose criterion for evaluating performance, since a detection is considered correct when the model roughly matches the correct answer and the prediction. mAP50 increases indicate that the model is generally able to accurately identify the target region and is a commonly used accuracy metric in object detection and segmentation. The mAP50-95 is the average accuracy calculated at each threshold value where the overlap threshold between the correct and predictive regions varies from 50% to 95% in 5% increments, so it has a more stringent performance evaluation than mAP50. The increase in mAP50-95's indicates that the model is not only detecting the target area but also more accurately predicting the contour and shape of the target. Therefore, mAP50-95 is considered an important indicator in the overall performance evaluation of the model. It was confirmed that the seg_loss of both Train data and Validation data decreased during the training process of this model. This indicates that the model has not only improved its segmentation ability appropriately for the training data but has also demonstrated high performance for unknown Validation data. The increase in both mAP50 and mAP50-95 for the Validation data indicates that the model's detection accuracy has consistently improved across a range of criteria, from a moderate threshold of 50% overlap between the correct and predicted regions to a more stringent range of 50% to 95% overlap.

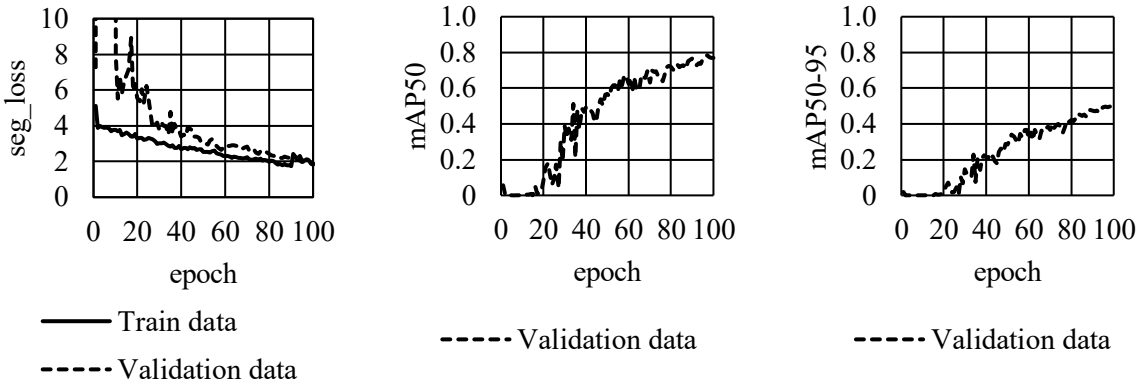
This result suggests that the model is not over-trained and that learning is progressing while maintaining good generalization performance. Interestingly, the 320x320 image size yielded higher training accuracy compared to 640x640 and 1280x1280 resolutions. This can be attributed to several factors:

Computational efficiency: Lower resolution images (320x320) are less computationally expensive, allowing for more training iterations per epoch. This enables parameter updates on a larger volume of data within the same training time.



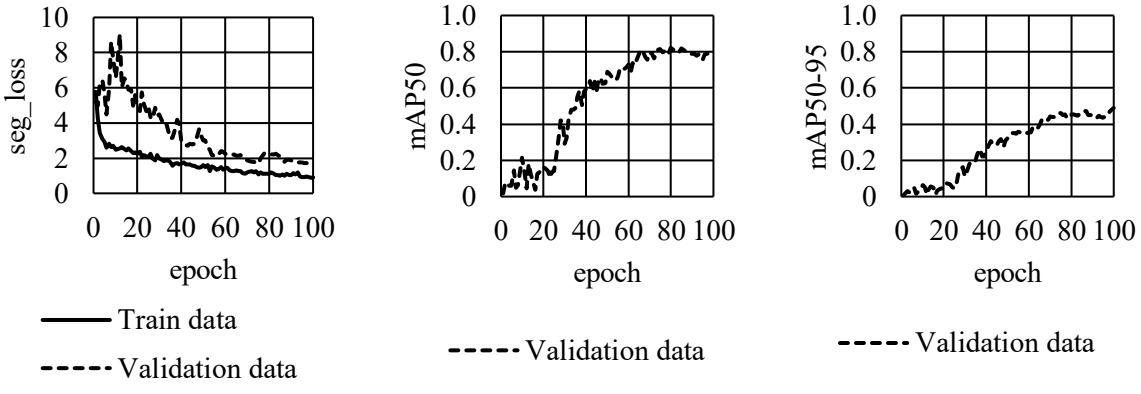
(a) seg_loss (b) mAP50 (c) mAP50-95

Figure 3. The training status of YOLOv8-seg (320x320)



(a) seg_loss (b) mAP50 (c) mAP50-95

Figure 4. The training status of YOLOv8-seg (640x640)



(a) seg_loss (b) mAP50 (c) mAP50-95

Figure 5. The training status of YOLOv8-seg (1280x1280)

Focus on macroscopic features: The lower resolution may have encouraged the model to concentrate on broader, more general features of the objects rather than relying on excessive micro-details. This approach likely contributed to better overall performance.

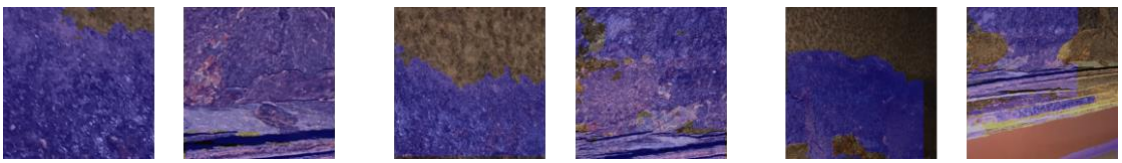
Balanced learning: For 640x640 and 1280x1280 images, the increased computational load resulted in fewer training cycles per epoch, potentially leading to insufficient learning.

Complexity of high-resolution learning: The increase in resolution necessitated learning more intricate details, which complicated the learning process and heightened the risk of over-learning and overreliance on local features.

These factors collectively contributed to the decreased accuracy observed in the 640x640 and 1280x1280 cases compared to the 320x320 resolution. The results underscore the importance of balancing image resolution, computational resources, and the nature of the features being learned in object detection tasks.

4. EVALUATION OF SYSTEM PERFORMANCE

In this study, unused test data was input to the trained YOLOv8-seg model to evaluate its ability to properly extract abnormal rust regions. The test data used were surface images of non-uniformly corroded weathering steel bridges in a real bridge environment. Figure 6 shows an example of detection results when an image of the same size as the training data was inputted, and Figure 7 shows detection results when the original image was input as test data. In the detection of abnormal rust in the cropped image, many areas were accurately captured in pixel units for both image sizes. However, as the image size was increased, the detection tended to miss dark-colored rusts when adjacent rusts in the same abnormal rust area were of different colors. There were also cases of mistakenly missing areas where corrosion was in progress, but the rust looked normal. On the other hand, when the original image was inputted, there was a tendency for even non-abnormal areas to be extracted, although the number of cases of missing abnormal rust was small. This tendency was particularly pronounced for models trained on 320x320 pixel images.



(a) 320x320 (b) 640x640 (c) 1280x1280

Figure 6. Examples of detection results testing small area image

(a) 320x320 (b) 640x640 (c) 1280x1280

Figure 7. Detection results testing the original image

As a quantitative evaluation of the system, this study was also validated using IoU (Intersection over Union), which indicates the overlap between the correct and predicted regions. The formula for calculating IoU is shown in Equation (1).

$$IoU = TP / (TP + FN + FP) \tag{1}$$

where TP (True Positive) is defined as correctly detected regions, FP (False Positive) as incorrectly detected regions, and FN (False Negative) as missed regions. Table 3 shows the

average values of IoU for the test data of the systems constructed for each image size. For example, an IoU of 0.65 indicates that the correct and predicted regions are about 1/9 of the way off in horizontal and vertical directions. The mIoU of the system constructed in this study ranged from 0.7356 to 0.9173 for the cropped test data and from 0.6195 to 0.6940 for the original image. These results indicate that the system can extract abnormal rust regions with a certain degree of accuracy.

Table 3. MIoU for system test data

	mIoU	
	Small Area	Original Image
320x320pixels	0.9173	0.6195
640x640pixels	0.7356	0.6940
1280x1280pixels	0.7371	0.6755

5. CONSIDERATION

This study compared the performance of models trained on images with different resolutions (320x320, 640x640, and 1280x1280) to evaluate their ability to extract abnormal rust regions from cropped images and original images used as test data.

The results showed that models trained on 640x640 and 1280x1280 pixels images tended to miss more test data and exhibited lower IoU values compared to the model trained on 320x320 pixels images when evaluated on cropped test images. This outcome is likely due to the increased computational load associated with high-resolution images, which reduces the number of parameter updates within the same training time, leading to insufficient training. Additionally, as image resolution increases, models tend to focus on learning fine-grained features, potentially at the expense of capturing the macro-level characteristics of rust anomalies. This imbalance may have contributed to the increased number of missed detections. Furthermore, in actual environments, abnormal rust areas often exhibit pronounced differences in color and texture. High-resolution models may become overly sensitive to these subtle variations, resulting in higher rates of false positives and missed detections.

When the original images were used as test data, excessive detections were observed in the model trained on 320x320 images, and the IoU tended to be lower than those of models trained on higher resolutions. This result can be attributed to the narrower field of view in 320x320 training images, which caused the model to focus heavily on features specific to abnormal rust. As a result, it struggled to distinguish between normal and abnormal rust areas, leading to over-detection by incorrectly identifying non-abnormal regions as rust anomalies. Since original images cover a broader area than cropped training images, the 320x320 model likely lacked the capacity to effectively analyze such wide-ranging information. Conversely, models trained on 640x640 and 1280x1280 resolutions were better equipped to learn higher-resolution features, enabling them to more accurately distinguish between normal and abnormal rust regions in original images. This capability appears to have helped suppress over-detection. To improve the accuracy of abnormal rust detection, it is crucial to train models using datasets that incorporate multiple image resolutions simultaneously. This approach would allow models to effectively capture both fine-grained and macro-level features. Additionally, employing data augmentation techniques such as scaling transformations, rotations, and adjustments to brightness and contrast could enhance the model's generalization performance. These strategies would help ensure stable detection accuracy under real-world conditions.

6. CONCLUDING REMARKS

This study developed a system utilizing YOLOv8-seg to extract abnormal rust areas classified as appearance grade 1 and 2 from the surface of non-uniformly corroded weathering steel bridges. The research also investigated the impact of different image resolutions used for training. The results showed that the detection accuracy of abnormal rust areas differs among models with different image resolutions. In particular, the model trained at 320x320 pixel resolution showed high detection accuracy for the cropped image, while the test data using the original image showed significant over-detection, and the IoU values tended to be lower than those of other resolution models. In contrast, the 640x640 and 1280x1280 pixel models were able to suppress over-detection for the original images, but some anomalous regions were missed for the cropped images.

Based on these results, future work should focus on achieving consistent detection accuracy for various image sizes through the introduction of multi-scale learning and further enhancement of data expansion. Another promising direction is the development of systems for advanced feature extraction of anomalous regions and detailed classification of the type and progression of anomalies. Furthermore, it is also essential to consider lightening the weight of the system in consideration of system operability and real-time performance under real-world conditions. With these improvements, this system is expected to contribute as a more accurate and reliable abnormal detection tool in bridge maintenance and management.

ACKNOWLEDGEMENTS

The authors thank Dr. Atsumi Imai of the Public Works Research Center for providing graded rust images of weathering steel and image data of actual bridges, as well as advice on rust evaluation of weathering steel bridges.

REFERENCES

Aoshima, K., Yamamoto, T., Nakano, S., and Nakamura, H. (2018), Study on Variant Extraction of Concrete Structures Using Image Recognition by Deep Learning, Journal of Japan Society of Civil Engineers, Ser.E2(Materials and Concrete Structures), 74(4), 293-305. (in Japanese)

Tabata, Y., Dang, J., Haruta, D., Shrestha, A., and Chun, P. (2018), Unmaned Inspection Orientated UAV Bridge Inspection and Detection Using Deep Learning, Journal of Japan Society of Civil Engineers, Ser.F4(Construction and Management), 74(2), I_62-I_74. (in Japanese)

Goto, S., Aso, T., and Miyamoto, A. (2006), A Rust Evaluation Method for Weathering Steels Based on Image Processing And Pattern Recognition, Journal of Japan Society of Civil Engineers, Ser.F, 62(4), 674-683. (in Japanese)

Imai, A., Yamamoto, T., and Aso, T. (2012), A Study of Partial Repair Painting for Weathering Steel Bridge, Journal of Japan Society of Civil Engineers, Ser.A1(Structural and Earthquake Engineering), 68(2), 347-355. (in Japanese)

Tai, M., Sekiya, H., Okatani, T., Nakamura, S., Shimizu, T. (2021), Effect of CNN Model and Image Size on the Accuracy of Rust Appearance Grading Discrimination of Weathering Steel Plates, Journal of AI and Data Science, Vol.2, J2, 378-385. (in Japanese)

Joseph Redmon, Santosh Divvala, Ross Girshick, and Ali Farhadi. (2016), You only look once: Unified, real-time object detection. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 779– 788.