Deep Learning-based Rail Surface Condition Evaluation

Shilin Hu[1], Ke Ma[2], Sagnik Das[1], Dichang Zhang[1], Dimitris Samaras[1]
[1]Stony Brook University [2]Snap Inc.
{shilhu, kemma, sadas, diczhang, samaras}@cs.stonybrook.edu

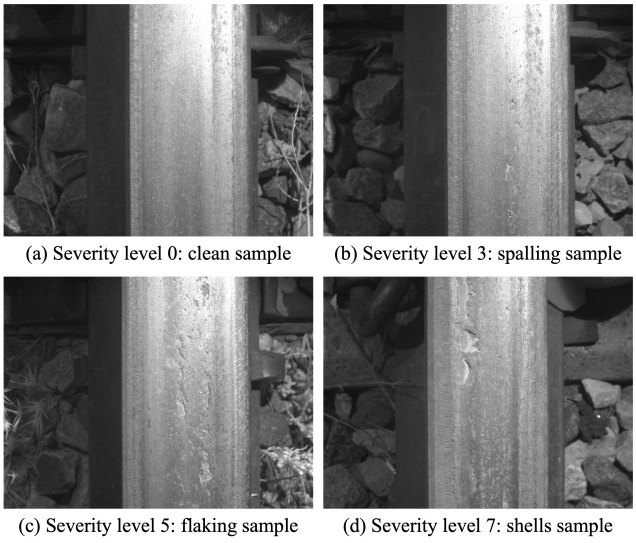
Abstract

Rail surface defects can develop into rail cracking, leading to rail failures that pose serious threats to transportation safety. Assessing the severity level of these surface defects is critical for proactive maintenance and long-term asset management. However, existing vision-based inspection systems primarily focus on classifying defect types or locating their positions on the rail surface, offering limited assistance for maintenance planning. To address this gap, we present a unified deep learning framework for automated rail surface defect severity evaluation. Our framework produces direct and interpretable evaluations of the overall surface condition, ranging from level 0 (no defect) to level 7 (severe defect), to support effective maintenance decision-making. The framework comprises three key components: (i) a segmentation module that identifies rail surfaces, eliminating interference from background pixels, (ii) an alignment module that standardizes the pose of rail surfaces to mitigate scale and rotation variance, and (iii) a classification module that predicts defect severity on the aligned rail surfaces. On a new benchmark of expert-labeled high-resolution images, our system achieves 81.8% main-diagonal and 95.9% tri-diagonal accuracy, processing up to 47 images per second. These results demonstrate the reliability and efficiency of our framework for large-scale rail surface monitoring. The dataset and code are available at https://github.com/cvlab-stonybrook/RailEval.

1. Introduction

Rail transportation plays a crucial role in modern society, and rail track maintenance is essential to ensure safety. According to the Federal Railroad Administration (FRA) Office of Safety database [8], in 2024, 1,707 train accidents were reported in the United States, with approximately 24% attributed to track-related causes, making them the second leading factor. These failures not only result in financial loss exceeding $315 million, but also threaten public safety.

Among the various track-related issues, surface-level rail defects are especially critical. Known as Rolling Contact Fatigue (RCF), these defects, such as spalling, flaking, and shells (as shown in Fig. 1), develop under repeated wheel-rail interactions and can accelerate the formation of subsurface cracks that lead to structural rail failures. Furthermore, these visible surface defects can obscure internal damage during non-visual inspections such as ultrasonic testing [1], underscoring the importance of timely and accurate surface condition evaluation.



(a) Severity level 0: clean sample (b) Severity level 3: spalling sample
(c) Severity level 5: flaking sample (d) Severity level 7: shells sample

Figure 1. Sample images of different rail surface severity levels. (a) shows a sample image of severity level 0 (no defect), (b) shows a spalling defect at severity level 3, (c) shows a flaking defect at severity level 5, and (d) presents shelling at severity level 7.

To support efficient and safe rail operations, maintenance planners and field inspectors need to conduct timely, objective assessments of rail surface condition. Traditional manual inspection methods are inherently subjective, and machine vision technology is used to augment these inspections by ensuring consistency across evaluations. While rail-bound inspection vehicles have enabled scalable vision-based rail surface monitoring, most current systems [14, 26, 27, 30, 31] focus on detecting defect types and locations, overlooking the severity of surface wear. This omission limits their utility for maintenance

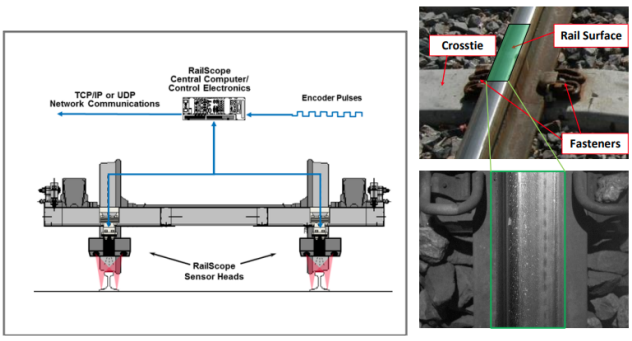


Figure 2. Image Acquisition System (Figure courtesy of [20]). The RailScope system [15] consists of two rail surface image sensor heads, one for each rail, and central computer/control electronics. Inside each sensor head is a laser light source that is used to illuminate the rail and a high-resolution digital camera to capture the rail surface images. As the vehicle travels down the track, high-resolution images of the rail surface are acquired.

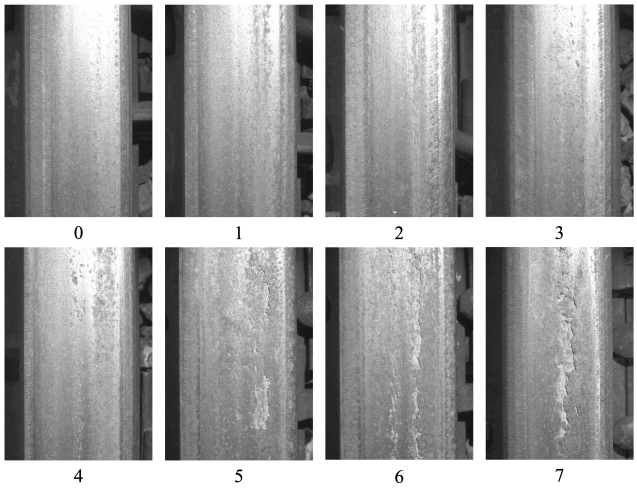


Figure 3. Examples of the 8 different severity levels. 0 means no defects, and 7 means the most serious defects.

planning, where severity levels are critical for prioritizing and scheduling track grinding, repair, or replacement [1].

In this paper, we present a deep learning-based framework for classifying surface defect severity to support maintenance planning and enhance rail transportation safety. Our method analyzes high-resolution rail images captured by the RailScope system [15], as shown in Fig. 2. Each rail surface image is categorized into one of eight severity levels (0 to 7), defined by domain experts to reflect operational maintenance thresholds. Fig. 3 illustrates examples from different severity levels: level 0 indicates no damage, level 4 shows moderate damage requiring maintenance, and level 7 signifies severe defects that necessitate track replacement.

Classifying defect severity from these images poses several challenges, including background interference and variations in the rotation, scale, and position of the rail surface within each frame. To address these issues, our framework comprises three core components: i) a **segmentation module** that predicts binary masks to extract the rail surface region and suppress background noise; ii) an **alignment module** that estimates a geometric transformation from the masked image to normalize the pose of the cropped rail surface, and applies it to the original input to preserve rail surface information and reduce errors from segmentation, and iii) a **classification module** that predicts the severity level from the normalized rail surface crop.

A further challenge lies in the limited availability of annotated real-world data, particularly for high-severity defects. Because rail defects are typically corrected through routine maintenance, severe cases are infrequent in standard operations. To address this, we introduce a new benchmark dataset collected at the Transportation Technology Center (TTC), consisting of 1,132 high-resolution (1200×1600) rail surface images captured from the High Tonnage Loop, where defects are intentionally allowed to develop for research purposes. All the images are labeled by field experts using an eight-level severity scale. The TTC benchmark dataset spans all severity levels and provides a valuable resource for future research in rail inspection and infrastructure monitoring.

We evaluate our proposed framework on the TTC benchmark dataset. Our method achieves a main-diagonal accuracy of 81.8% (exact severity match), a tri-diagonal accuracy of 95.9% (± 1 severity tolerance), and an inference speed of 47 frames per second (FPS) using batched inputs.

To enhance reliability, we incorporate an anomaly detection module designed to identify rail images that fall outside the defined eight-level severity scale. These outlier cases often feature under-represented visual patterns, such as rail joints, switches, or welds, which can lead to unreliable severity predictions. The module filters such images from automated processing, flagging them for human inspection. It achieves an accuracy of 99.78%. This filtering reduces the risk of misclassification in out-of-distribution (OOD) scenarios, without interfering with standard automated analysis. Overall, our framework combines high efficiency and accuracy to support reliable, low-latency, high-throughput vision-based inspection workflows, serving as an assistive tool for rail inspectors and maintenance planners in infrastructure monitoring and decision-making.

Our contributions are summarized as follows:

- A unified deep learning framework for rail surface defect severity classification, integrating segmentation, alignment, and classification modules, and achieving 81.8% accuracy at 47 FPS.
- A new benchmark dataset from the TTC with expert annotations covering all eight severity levels.
- An anomaly detection module for filtering out OOD cases

for human inspection, achieving 99.78% accuracy.

2. Related works

Track inspection is a longstanding challenge, involving evaluation of track components such as fasteners [9, 19], joint bars [2], and railroad ties [18, 23]. Rail defect evaluation is particularly crucial, as defects can arise from various factors like environmental conditions, train speed, and tonnage. Although Ultrasonic Nondestructive Testing (NDT) has been the standard for internal flaw inspection [1, 3, 32], it lacks sensitivity to surface-level degradation. The risk of surface defects propagating into subsurface cracks and causing rail failure highlights the need for inspection methods focusing on rail surface defects.

Two major technologies for rail surface defect evaluation are: 1) Eddy Current Inspection: This method uses alternating magnetic fields to induce eddy currents in the rail. Surface defects are detected when these currents are disrupted. However, this method is sensitive to variations in lift-off, requiring a consistent distance between the probe and the rail surface [22]. 2) Image-based Inspection: This method detects rail surface defects through analysis of rail images. After collecting an image dataset, machine learning algorithms are applied to identify defects in the rail surface. Early machine vision approaches include a filter-based system focused on corrugation, using Gabor filters to extract texture information [21]. Li *et al*. [17] developed a vision system that enhances images and uses a maximum entropy thresholding algorithm to detect defects. With the advancement of deep learning, vision-based methods have improved significantly. Shang *et al*. [26] introduced a two-stage model that locates rail surfaces and uses a convolutional neural network (CNN) for detecting defective rails. Faghih-Roohi *et al*. [7] used CNNs to classify various rail surface types. Chen *et al*. [6] employed a Faster RCNN architecture for defect detection and location. However, these systems primarily focus on defect detection or type classification and do not address severity level estimation, which is essential for maintenance prioritization. Comprehensive surveys of surface inspection methods can be found in [10, 16].

Ma *et al*. [20] were the first to emphasize the importance of rail surface severity-level evaluation, but the lack of code and data hinders reproducibility. They used the Generalized Hough transform and edge maps from a trained structured random forest to segment rail surfaces and a stacked ensemble of Support Vector Machine (SVM) classifiers for severity classification. However, their system suffers from high latency, over 9 seconds per image. In contrast, our deep learning framework achieves significantly faster inference while improving classification accuracy. The framework is designed for assistive deployment, supporting fast, automated evaluation while allowing human oversight in OOD cases, aligning with safety-critical inspection goals.

3. Method

3.1. Overview

We propose a unified framework for classifying the severity of rail surface defects to support condition-based maintenance planning. The system operates on high-resolution rail surface images collected by the RailScope platform [15] and comprises three key modules: a segmentation module that isolates the rail surface region, removing irrelevant background content; an alignment module that estimates an affine transformation to normalize the pose and scale of the rail surface, addressing variations caused by motion and viewpoint; and a classification module that predicts a severity level from 0 to 7 based on the aligned crop. The full pipeline is illustrated in Fig. 4. Given a captured image, the segmentation module first predicts a binary mask to locate the rail surface. This masked region is then processed by the alignment module, which estimates a transformation matrix to center and standardize the geometry of the rail segment. The aligned crop is finally passed to the classification module, which outputs a defect severity prediction.

The segmentation and alignment modules are pretrained on a separate proprietary dataset and remain fixed throughout training and evaluation on our benchmark. Their architecture and integration are essential for enabling reliable rail surface location and geometric normalization under real-world operating conditions.

An additional anomaly detection module is included to identify and filter out rail images that fall outside the defined severity scale, such as those containing rail joints, welds, or switches. These outliers are excluded from automated severity classification and flagged for human review, adding an extra layer of reliability to the system.

3.2. Segmentation module

To suppress background interference in assessing the rail surface defect severity levels in the captured rail images, we employ a segmentation module that isolates the rail surface region. Unlike the prior approaches based on edge detection [20, 30], we formulate this as a binary semantic segmentation task that classifies each pixel as either rail surface (1) or background (0). The module adopts a lightweight encoder-decoder structure with a MobileNetV2 backbone [24] and an ASPP-based decoder [4], with skip connections to preserve spatial details. Input images are resized to 512×512 before segmentation.

Its output, a binary mask, is used to extract the rail surface via element-wise multiplication, removing background content before alignment. This focuses the alignment module on rail geometry alone, reducing distractions from background noise. Segmenting first also decouples background variation from pose normalization, resulting in more stable and accurate transformations.

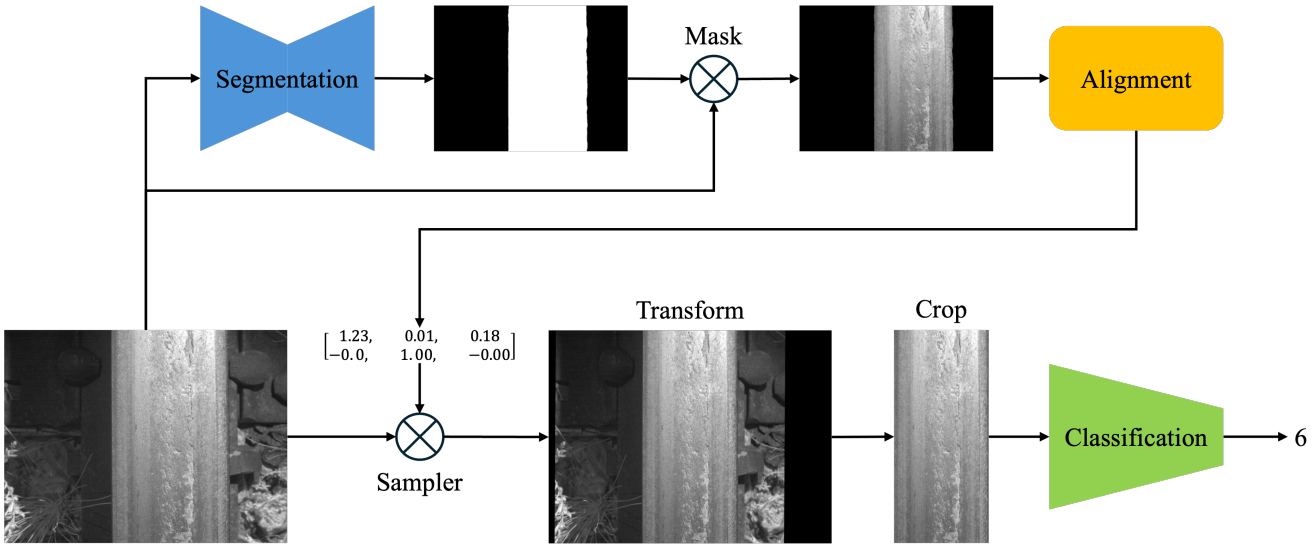


Figure 4. Illustration of the pipeline. The model takes the captured image as input. The segmentation module predicts the rail surface mask, then the alignment module predicts the transformation matrix and applies it to the input to correct the pose of the rail surface, and finally, the classification module outputs the prediction of the rail track surface defect severity level.

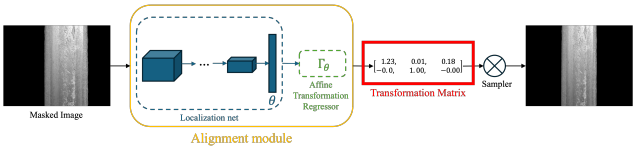


Figure 5. The architecture of the standalone alignment module. Modified from the spatial transformer [13], the input of the masked rail surface is passed to a location net that predicts transformation parameters θ. These parameters are then regressed to an affine transformation matrix. Finally, the affine transformation is applied to the input and produces the corrected rail surface.

3.3. Alignment module

Due to vibrations and motion from the rail-bound vehicles, captured rail surface segments often appear at varying positions, orientations, and scales. To normalize these variations, we incorporate an alignment module based on the spatial transformer network (STN) framework [13]. This module takes a masked rail surface image as input and predicts a 2×3 affine transformation matrix, which is then applied to reposition the rail surface to the center of the image frame (see Fig. 5).

To preserve detail, the alignment module operates on the full-resolution image (1200×1600). The segmentation mask is upsampled to this resolution and used to isolate the rail surface. The resulting masked image is passed to the alignment module, which estimates the transformation matrix. This transformation is applied to the original image to produce a standardized, centered view of the rail surface segment for downstream classification.

3.4. Classification module

Following segmentation and alignment, we obtain a pose-normalized image in which the rail surface is centered and scaled to a fixed width of 535 pixels, corresponding to the dataset-wide average. From this standardized representation, we resize it to 448×448 pixels before feeding it into the classification module to assess surface condition.

The classification module predicts one of the eight severity levels (0–7) based on the visual characteristics of the aligned rail segment. It is trained using the Cross-Entropy (CE) loss between the predicted severity score and the ground truth label:

$$L_{cls} = CE(\text{pred}, \text{label}) \tag{1}$$

We use ResNet18 [12] as the backbone of our classification module, selected for its balance of accuracy, efficiency, and suitability for deployment.

3.5. Anomaly detection module

While the classification module is trained to assign rail surface segments to one of eight severity levels (0–7), certain inputs fall outside this predefined taxonomy, such as welds, joints, and switches, which are not represented in the training data. These OOD cases exhibit geometry or texture patterns that differ significantly from the benchmark distribution, where a single, clearly visible rail surface is consistently present. If not explicitly handled, such inputs can lead to unreliable predictions. Fig. 6 shows representative examples of anomalous rail images.

We formulate anomaly detection as a binary classification task, aiming to distinguish between in-distribution

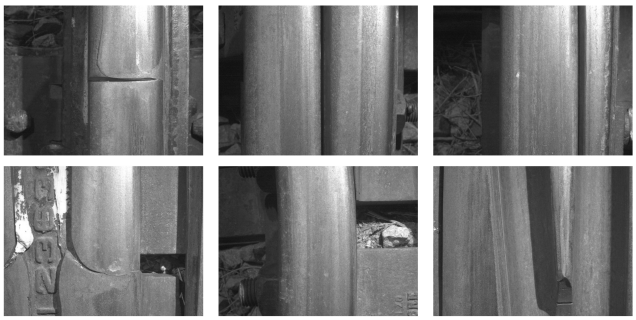


Figure 6. Examples of out-of-distribution rail surface inputs flagged by our anomaly detection module. These include welds, joints, and switches, which are excluded from severity classification and routed for manual inspection.

rail surface images and anomalous inputs exhibiting rare or structurally distinct patterns. The module is trained using benchmark images as normal examples and a curated set of anomaly images as outliers.

Although the benchmark dataset contains only clean, in-distribution images, the anomaly detection module is intended for real-world field inspection. It operates alongside the main classification pipeline to flag inputs that deviate from the expected distribution. These flagged cases are then routed for manual review to ensure reliable handling of atypical or ambiguous inputs.

4. Experiments & Results

Datasets. We conduct experiments on two datasets: a **benchmark dataset** for severity classification and a curated **anomaly dataset** for out-of-distribution detection. The rail surface images are collected using the RailScope system [15], a state-of-the-art image acquisition platform that captures top-down views of the rail head in real-time. Equipped with a laser light source and a high-resolution CCD digital camera, the system produces images with sufficient detail to identify pitting, spalling, and surface cracking. The benchmark dataset contains 1,132 rail images. Each image is labeled by domain experts using an eight-level severity scale, from level 0 (no visible defect) to level 7 (severe degradation). The labeling criteria are provided in the supplementary material. The distribution of severity levels is shown in Tab. 1. In addition, we construct an anomaly dataset containing 181 OOD images, such as rail joints, switches, and welds. These images do not conform to the geometry or appearance of standard rail surfaces and are used to train and evaluate the performance of our anomaly detection module. All results are reported using 4-fold cross-validation with stratified splits to maintain class proportions across equally sized subsets.

Evaluation metrics. For severity classification, we report both *main-diagonal accuracy* and *tri-diagonal accuracy*, following Ma *et al*. [20]. The main-diagonal accuracy measures exact matches between the predicted and the ground truth severity levels. The tri-diagonal accuracy allows for the predictions within ± 1 of the ground truth label, reflecting practical tolerance during field assessments where severity may transition gradually due to grinding or annotation ambiguity. For anomaly detection, we report a standard classification accuracy to reflect the module's ability to distinguish OOD cases from normal rail surfaces.

Table 1. Image counts per severity level in the benchmark dataset.

Severity Level	0	1	2	3	4	5	6	7
Image Count	118	160	91	163	66	277	167	90

All experiments are conducted on a platform equipped with an Intel Xeon Gold 5218 CPU and a single NVIDIA TITAN RTX GPU. Our implementation is based on PyTorch. For training the classification module, we use the Adam optimizer with a learning rate of $1\mathrm{e}{-4}$. The segmentation and alignment modules are pretrained on a proprietary dataset and remain fixed during all experiments on the benchmark. Additional details and evaluations of these modules are provided in the supplementary material.

4.1. Segmentation and Alignment Results

The segmentation and alignment modules are pretrained on a proprietary dataset and applied directly to the benchmark without additional fine-tuning. This zero-shot transfer works effectively because of the consistent geometric structure of the rail surfaces—specifically, the approximately parallel boundaries of the rail head—which facilitates generalization across severity levels and imaging conditions.

As illustrated in Fig. 7, we present qualitative examples from severity levels 1, 3, 5, and 7. The first column shows the original rail surface images. The second column displays the predicted binary segmentation masks. These masks generally exhibit high-quality delineation of the rail surface, although minor imperfections, such as small holes or missing sections, can be observed (in rows 1 and 3).

The third column presents the output of the alignment module, where the predicted affine transformation matrix is applied directly to the original image. This design helps avoid propagating segmentation errors: applying the predicted transformation to the original image, rather than the masked one (as in Fig. 5), prevents small imperfections, such as holes or boundary artifacts, from being carried into the aligned output. As a result, the output remains stable and visually consistent, preserving fine-grained details and surface continuity. The fourth column overlays a standardized target region (in red) on each aligned image to illustrate the effectiveness of our design. The rail surfaces are consistently centered and cropped at a standardized scale and position across the benchmark dataset. This normalization

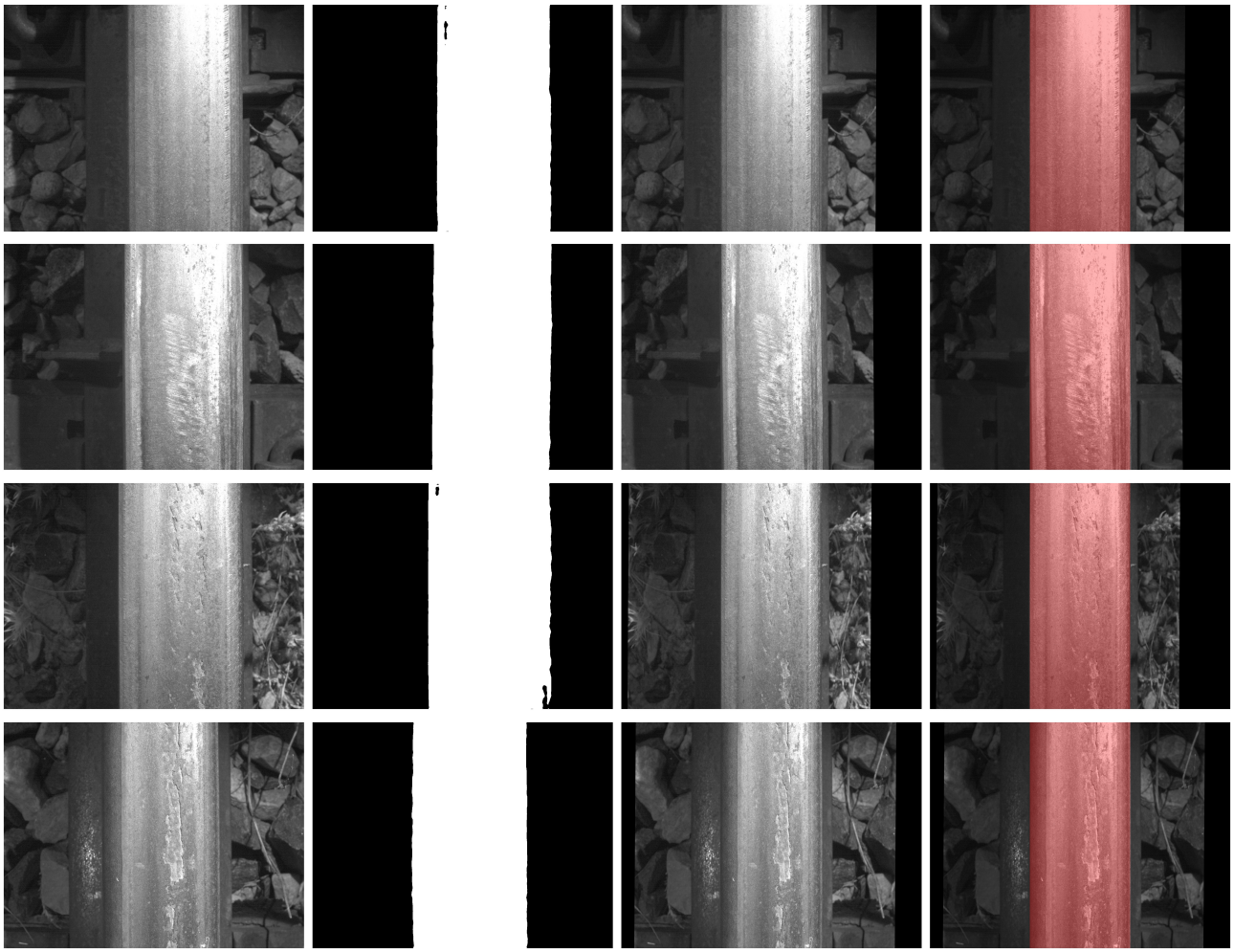


Figure 7. Qualitative results of the segmentation and alignment modules. From left to right: input rail image, predicted binary segmentation mask, output of the alignment module after spatial transformation, and overlay of the fixed target region (in red) on the aligned image. The proposed method successfully corrects the pose and scale of the rail surface while mitigating errors introduced by the segmentation mask.

reduces geometric variability in the input and enables the classification module to focus on surface condition differences rather than irrelevant pose or alignment variations.

In summary, the segmentation and alignment modules jointly produce clean, consistently framed rail surface crops. By tolerating minor segmentation imperfections and enforcing pose normalization, they ensure reliable, uniform inputs for accurate downstream severity classification.

4.2. Classification Results

Quantitative Results. We first compare our baseline framework (without augmentation) to a replicated implementation of Ma *et al*. [20], the only prior work addressing rail surface defect severity classification. As shown in the first two rows of Tab. 2, our method achieves substantially better performance, with higher main-diagonal (79.1% vs. 71.7%) and tri-diagonal accuracy (95.3% vs. 92.3%), while also running over 400 times faster (47 FPS vs. 0.1 FPS). This improvement is attributed to the greater learning capacity of deep neural networks and the use of GPU acceleration, in contrast to the previous CPU-bound pipeline of random forests and SVMs.

Table 2. Quantitative results: We report the main-diagonal and tri-diagonal accuracy of our proposed framework, and inference speed (FPS). $*$ denotes results replicated by us based on [20].

Method	Main-diag Acc (%)	Tri-diag Acc (%)	Infer. speed (FPS)
Ma *et al*. [20]	71.7*	92.3*	0.1*
Ours w/o Aug	79.1	95.3	47
Ours w/ Image-level Aug	81.2	95.6	47
Ours w/ Pixel-level Aug	81.8	95.9	47

We then assess the effect of different data augmentation strategies, as shown in the last three rows of Tab. 2. Pixel-level augmentations, including brightness/contrast adjust-

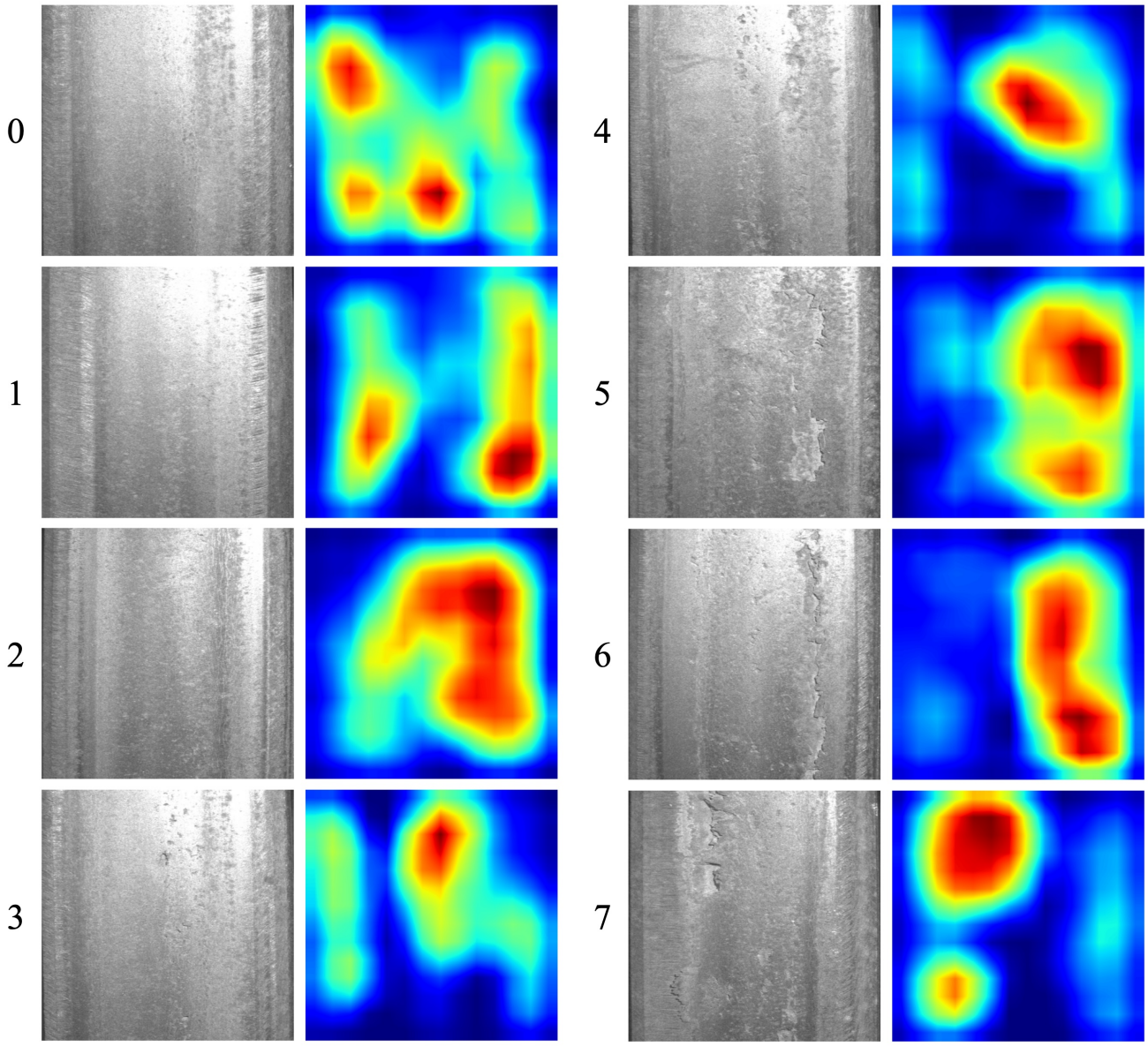


Figure 8. Qualitative results of the classification module. We show example images and corresponding Grad-CAM activations [25] for all eight severity levels. For minor defects (1, 2), the model attends to subtle surface scratches, often near the gauge side. For more severe cases (6, 7), attention focuses directly on spalling or flaking regions, indicating the model uses meaningful visual cues for severity prediction.

ment, ISO noise, Gaussian blur, and JPEG compression, result in the best performance, achieving 81.8% main-diagonal and 95.9% tri-diagonal accuracy. These augmentations enhance model robustness by introducing variation in visual conditions without distorting the structural characteristics of the defects. In contrast, mixed sample augmentation methods (*e.g.*, CutMix[28], Mixup[29], FMix[11], and GridMask[5]) lead to inferior results, as they often distort the rail surface or create unrealistic texture combinations that reduce label consistency and prediction reliability.

We evaluate the prediction performance of the best configuration using the confusion matrix shown in Fig. 9. The matrix shows a strong concentration of predictions along the main diagonal, indicating reliable per-class classification. Most off-diagonal errors occur within ± 1 severity level, particularly between levels 0–1, 3–4, and 5–6, reflecting the gradual and often ambiguous nature of surface degradation. Misclassifications beyond this range are rare and mainly between levels 1 and 3, which suggests that mild severity levels may share similar visual features.

Qualitative Results. We visualize the classification module's prediction process using Grad-CAM [25], as shown in Fig. 8. Representative examples from all eight severity levels are presented alongside their corresponding ac-

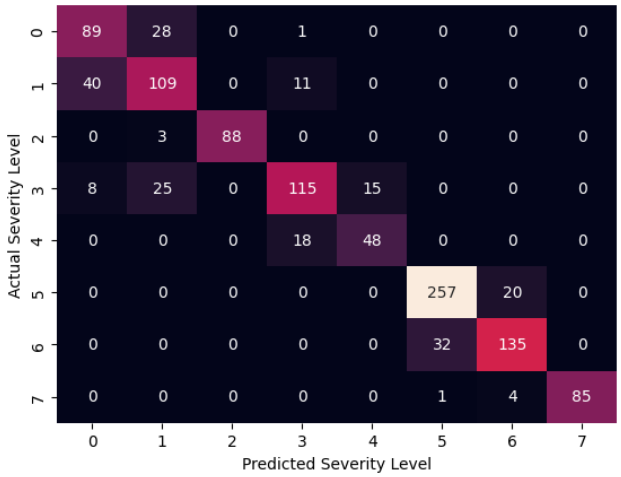


Figure 9. Confusion matrix of predicted versus actual rail surface severity levels. Our method shows strong diagonal alignment.

tivation maps. For level 0 (no defect), activations are diffused across the rail surface, indicating the absence of a specific defect region influencing the model's prediction. At lower severity levels (1–2), where defects typically appear as subtle scratches along the gauge side, the model's attention is appropriately located around these fine-grained features. As severity increases (levels 3–7), defects become more pronounced and visually diverse, manifesting as dents, spalling, or flaking. In these cases, the model consistently focuses on the precise defect regions regardless of their type or position, demonstrating that its predictions are guided by relevant visual cues. This behavior suggests that the model is not overfitting to specific defect types or spatial patterns, but instead generalizes well to a wide range of damage appearances.

These observations highlight that the classification module integrates diverse visual cues into a holistic assessment of rail surface condition. Instead of relying on fixed patterns or isolated features, it adapts to varying forms of surface damage. This supports the interpretability of the predicted severity level and makes it directly applicable to maintenance decision-making.

4.3. Anomaly Detection Results

For the anomaly detection module, we define the 1,132 images from the benchmark dataset as normal, and a curated set of 181 out-of-distribution images, captured at rail joints, welds, and switches, as anomalous. This is formulated as a binary classification task. A ResNet18 model is used as the backbone, achieving a classification accuracy of 99.78%. This result indicates a clear visual distinction between standard rail segments and under-represented rail configurations. The integration of the anomaly detection module enhances the robustness and deployment readiness of the proposed framework for real-world applications.

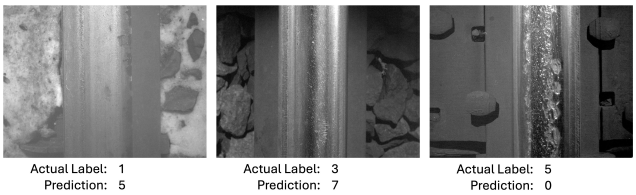


Figure 10. Failure cases on rail images with profiles differing from the benchmark. Environmental and operational variations in active service result in rail profiles that the framework fails to recognize.

5. Limitations

While our framework effectively evaluates rail surface condition on the proposed benchmark dataset, there are certain limitations to consider. First, the benchmark dataset was collected exclusively at the High Tonnage Loop at TTC over a limited time, resulting in relatively consistent rail profiles. However, rail conditions in active service can exhibit greater variability due to environmental and operational differences. As shown in Fig. 10, our model encounters failures when processing rail profiles that deviate substantially from those in the benchmark. Second, although our framework achieves 47 FPS, far exceeding the prior art, it remains below the 60 FPS target typically desired for full real-time performance in high-speed inspection scenarios. Future optimization may help close this gap for seamless integration into onboard or embedded inspection systems.

6. Conclusion

We presented a unified deep learning framework for vision-based rail surface inspection, combining segmentation, alignment, and classification modules. The system achieves accurate and efficient severity prediction on high-resolution rail images and incorporates an anomaly detection module to flag out-of-distribution inputs for human review. Evaluated on a newly introduced TTC benchmark dataset, the framework demonstrates strong classification performance and fast inference speeds, making it suitable for assistive deployment in industrial rail monitoring and maintenance workflows. To further enhance generalization under diverse operating conditions, future work will explore generative or synthetic data techniques to enrich training diversity and improve robustness in field deployments.

Acknowledgments

We thank Daniel Magnus from KLD Labs for leading data collection and providing industrial guidance throughout the project. We also thank Sean Woods, formerly of ENSCO, for coordinating data collection activities. We are grateful to Jay Baillargeon from the Federal Railroad Administration (FRA) for valuable feedback and continued support. This work was funded by the FRA and a gift from Nvidia Corp.

References

[1] FRA (Federal Railroad Administration). Track inspector rail defect reference manual, 2015. 1, 2, 3

[2] Andie Berry, Boris Nejikovsky, X Gilbert, and Ali Tajaddini. High speed video inspection of joint bars using advanced image collection and processing techniques. In *Proc. of world congress on railway research*, pages 619–622, 2008. 3

[3] Davide Bombarda, Giorgio Matteo Vitetta, and Giovanni Ferrante. Rail diagnostics based on ultrasonic guided waves: an overview. *Applied Sciences*, 11(3):1071, 2021. 3

[4] Liang-Chieh Chen, Yukun Zhu, George Papandreou, Florian Schroff, and Hartwig Adam. Encoder-decoder with atrous separable convolution for semantic image segmentation. In *Proceedings of the European conference on computer vision (ECCV)*, pages 801–818, 2018. 3

[5] Pengguang Chen, Shu Liu, Hengshuang Zhao, and Jiaya Jia. Gridmask data augmentation, 2020. 7

[6] Xiaobo Chen and Huimin Zhang. Rail surface defects detection based on faster r-cnn. In *2020 International Conference on Artificial Intelligence and Electromechanical Automation (AIEA)*, pages 819–822, 2020. 3

[7] Shahrzad Faghih-Roohi, Siamak Hajizadeh, Alfredo Núñez, Robert Babuska, and Bart De Schutter. Deep convolutional neural networks for detection of rail surface defects. In *2016 International joint conference on neural networks (IJCNN)*, pages 2584–2589. IEEE, 2016. 3

[8] Federal Railroad Administration. Fra office of safety analysis – train accident data. https://data.transportation.gov/stories/s/t23t-ueeq, 2025. 1

[9] Hao Feng, Zhiguo Jiang, Fengying Xie, Ping Yang, Jun Shi, and Long Chen. Automatic fastener classification and defect detection in vision-based railway inspection systems. *IEEE Transactions on Instrumentation and Measurement*, 63(4): 877–888, 2014. 3

[10] Wendong Gong, Muhammad Firdaus Akbar, Ghassan Nihad Jawad, Mohamed Fauzi Packeer Mohamed, and Mohd Nadhir Ab Wahab. Nondestructive testing technologies for rail inspection: a review. *Coatings*, 12(11):1790, 2022. 3

[11] Ethan Harris, Antonia Marcu, Matthew Painter, Mahesan Niranjan, Adam Prügel-Bennett, and Jonathon Hare. Fmix: Enhancing mixed sample data augmentation, 2021. 7

[12] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016. 4

[13] Max Jaderberg, Karen Simonyan, Andrew Zisserman, et al. Spatial transformer networks. *Advances in neural information processing systems*, 28, 2015. 4

[14] Xiating Jin, Yaonan Wang, Hui Zhang, Hang Zhong, Li Liu, QM Jonathan Wu, and Yimin Yang. Dm-ris: Deep multimodel rail inspection system with improved mrf-gmm and cnn. *IEEE Transactions on Instrumentation and Measurement*, 69(4):1051–1065, 2019. 1

[15] KLD Labs Inc. Rail surface assessment — railscope™ system. https://www.kldlabs.com/track-assessment-2/rail-surface-assessment/, 2025. 2, 3, 5

[16] Lei Kou. A review of research on detection and evaluation of the rail surface defects. *Acta Polytech. Hung*, 19:167–186, 2022. 3

[17] Qingyong Li and Shengwei Ren. A visual detection system for rail surface defects. *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews)*, 42(6): 1531–1542, 2012. 3

[18] Ying Li, Hoang Trinh, Norman Haas, Charles Otto, and Sharath Pankanti. Rail component detection, optimization, and assessment for automatic rail track inspection. *IEEE Transactions on Intelligent Transportation Systems*, 15(2): 760–770, 2013. 3

[19] Junbo Liu, Yaping Huang, Qi Zou, Mei Tian, Shengchun Wang, Xinxin Zhao, Peng Dai, and Shengwei Ren. Learning visual similarity for inspecting defective railway fasteners. *IEEE Sensors Journal*, 19(16):6844–6857, 2019. 3

[20] Ke Ma, Tomás F Yago Vicente, Dimitris Samaras, Michael Petrucci, and Daniel L Magnus. Texture classification for rail surface condition evaluation. In *2016 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pages 1–9. IEEE, 2016. 2, 3, 5, 6

[21] Clelia Mandriota, Massimiliano Nitti, Nicola Ancona, Ettore Stella, and Arcangelo Distante. Filter-based feature selection for rail defect detection. *Machine Vision and Applications*, 15:179–185, 2004. 3

[22] B Marchand, JM Decitre, and O Casula. Recent developments of multi-elements eddy current probes. In *Proceedings of the 17th World Conference on Nondestructive Testing, Shanghai, China*, pages 25–28, 2008. 3

[23] Alessandro Sabato and Christopher Niezrecki. Feasibility of digital image correlation for railroad tie inspection and ballast support assessment. *Measurement*, 103:93–105, 2017. 3

[24] Mark Sandler, Andrew Howard, Menglong Zhu, Andrey Zhmoginov, and Liang-Chieh Chen. Mobilenetv2: Inverted residuals and linear bottlenecks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4510–4520, 2018. 3

[25] Ramprasaath R Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *Proceedings of the IEEE international conference on computer vision*, pages 618–626, 2017. 7

[26] Lidan Shang, Qiushi Yang, Jianing Wang, Shubin Li, and Weimin Lei. Detection of rail surface defects based on cnn image recognition and classification. In *2018 20th International Conference on Advanced Communication Technology (ICACT)*, pages 45–51, 2018. 1, 3

[27] Biao Yue, Yangping Wang, Yongzhi Min, Zhenhai Zhang, Wenrun Wang, and Jiu Yong. Rail surface defect recognition method based on adaboost multi-classifier combination. In *2019 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)*, pages 391–396. IEEE, 2019. 1

[28] Sangdoo Yun, Dongyoon Han, Seong Joon Oh, Sanghyuk Chun, Junsuk Choe, and Youngjoon Yoo. Cutmix: Regularization strategy to train strong classifiers with localizable features, 2019. 7

[29] Hongyi Zhang, Moustapha Cisse, Yann N. Dauphin, and David Lopez-Paz. mixup: Beyond empirical risk minimization, 2018. 7

[30] Hui Zhang, Xiating Jin, QM Jonathan Wu, Yaonan Wang, Zhendong He, and Yimin Yang. Automatic visual detection system of railway surface defects with curvature filter and improved gaussian mixture model. *IEEE Transactions on Instrumentation and Measurement*, 67(7):1593–1608, 2018. 1, 3

[31] Hui Zhang, Yanan Song, Yurong Chen, Hang Zhong, Li Liu, Yaonan Wang, Thangarajah Akilan, and QM Jonathan Wu. Mrsdi-cnn: Multi-model rail surface defect inspection system based on convolutional neural networks. *IEEE Transactions on Intelligent Transportation Systems*, 23(8):11162–11177, 2021. 1

[32] G Zumpano and Michele Meo. A new damage detection technique based on wave propagation for rails. *International journal of solids and structures*, 43(5):1023–1046, 2006. 3