Rail-5k: a Real-World Dataset for Rail Surface Defects Detection

Zihao Zhang
The Key Laboratory of Road and Traffic Engineering, Ministry of Education
Shanghai Key Laboratory of Rail Infrastructure Durability and System Safety
Tongji University, Shanghai
1831412@tongji.edu.cn

Shaozuo Yu
Department of Computer Science
Tongji University, Shanghai
yushaozuo@tongji.edu.cn

Siwei Yang
Department of Computer Science
Tongji University, Shanghai
swyang.ac@gmail.com

Bingchen Zhao
Department of Computer Science
Tongji University, Shanghai
zhaobc.gm@gmail.com

Yu Zhou
The Key Laboratory of Road and Traffic Engineering, Ministry of Education
Shanghai Key Laboratory of Rail Infrastructure Durability and System Safety
Tongji University, Shanghai
wqhuo2785@163.com

Abstract

This paper presents the Rail-5k dataset for benchmarking the performance of visual algorithms in a real world application scenario, namely the rail surface defects detection task. We collected over 5k high quality images from railways across China, and annotated 1100 images with the help from railway experts to identify the most common 13 types of rail defects. The dataset can be used for two settings both with unique challenges, the first is the fully-supervised setting using the 1k+ labeled images for training, fine-grained nature and long-tailed distribution of defect classes makes it hard for visual algorithms to tackle. The second is the semi-supervised learning setting facilitated by the 4k unlabeled images, these 4k images are uncurated containing possible image corruptions and domain shift with the labeled images, which can not be easily tackle by previous semi-supervised learning methods. We believe our dataset could be a valuable benchmark for evaluating robustness and reliability of visual algorithms.

*Corresponding author

Table 1: Dataset compare

<table>
<thead>
<tr>
<th>Domain</th>
<th>Dataset</th>
<th>Task</th>
<th># class</th>
<th># image</th>
<th># box per image</th>
<th>Resolution</th>
<th>Annotation Quality</th>
</tr>
</thead>
<tbody>
<tr>
<td>Rail Defects</td>
<td>Delft [4]</td>
<td>cls</td>
<td>6</td>
<td>3240</td>
<td>1</td>
<td>$10^4$ gray-scale</td>
<td>image-level</td>
</tr>
<tr>
<td></td>
<td>RSDDs [6]</td>
<td>seg</td>
<td>2</td>
<td>195</td>
<td>5</td>
<td>$10^4$ gray-scale</td>
<td>image-level</td>
</tr>
<tr>
<td></td>
<td>CRRC [5]</td>
<td>det</td>
<td>3</td>
<td>&gt;1000</td>
<td>1</td>
<td>$10^4$ gray-scale</td>
<td>band-level</td>
</tr>
<tr>
<td>Rail-5k(labeled)</td>
<td></td>
<td>det</td>
<td>13</td>
<td>1100</td>
<td>22.9</td>
<td>$10^7$ RGB</td>
<td>instance-level</td>
</tr>
<tr>
<td>Natural Image</td>
<td>VOC-2007</td>
<td>det</td>
<td>20</td>
<td>12974</td>
<td>3.1</td>
<td>-</td>
<td>instance-level</td>
</tr>
<tr>
<td></td>
<td>VOC-2012 [3]</td>
<td>det</td>
<td>20</td>
<td>34071</td>
<td>2.7</td>
<td>469 x 387 RGB</td>
<td>instance-level</td>
</tr>
<tr>
<td></td>
<td>ILSVRC-2014 [2]</td>
<td>det</td>
<td>200</td>
<td>516840</td>
<td>1.1</td>
<td>482 x 415 RGB</td>
<td>instance-level</td>
</tr>
<tr>
<td></td>
<td>MS COCO 2018 [16]</td>
<td>det</td>
<td>80</td>
<td>163957</td>
<td>7.3</td>
<td>-</td>
<td>instance-level</td>
</tr>
<tr>
<td></td>
<td>OID V6 [14]</td>
<td>det</td>
<td>600</td>
<td>1910098</td>
<td>8.4</td>
<td>-</td>
<td>instance-level</td>
</tr>
</tbody>
</table>

1 Introduction

The introduction of large scale annotated datasets such as ImageNet [2] greatly speeds up the development of deep-learning based vision algorithms [9]. Deep learning algorithms pre-trained on ImageNet [2] has also been shown to effectively transfer between domain and tasks such as object detection [21] or medical image analysis [12].

As an important basic infrastructure of human life, the maintenance and status analysis of railways has a real world economy and safety-focused value. However, current datasets in the railways domain are either limited in size [6], quality of images [4, 5, 6], or the annotation types [5, 6]. The limited size and quality of currently available dataset are not yet ready for support the training of deep learning methods.

Our dataset has enough high quality images captured from real-world railway to enable the training of deep learning models. Besides the labeled set with 1.1k images, we also provide a unlabeled set of 4k images to enable a semi-supervised setting. Several unique characteristics of our dataset also poses new challenges to vision algorithm. The first challenge is the long-tailed distribution of classes presented in out dataset, the imbalance ratio of the most majority class to the most minority class is up to 40.98, it has been shown that the long-tailed distribution would greatly hurt the performance of the learned model [17, 8]. Besides the long-tailed class distribution in the labeled set of the dataset, the unlabeled set of images also poses a difficult scenario of semi-supervised defect detection, semi-supervised object detection is a relatively new task with few recent works [7, 22], the previous method often assumes that the unlabeled set is also curated. However, in our case, the unlabeled set is uncurated with multiple unknown image corruptions and unseen object in the labeled set. Given these unique properties, we believe that our proposed dataset could not only facilitate the development of algorithms for rail surface defects detection, but also the development for a more robust vision model to handle the long-tailed distribution and possible corruptions in the unlabeled set.

2 Related work

Traditional inspection methods like subjective manual observation, sampling checking, are all qualitative or compensating methods, can not provide a digital and automatic decision-making basis for intelligent maintenance of the whole line. Our dataset mainly focus on the task of defects detection, we summarize relevant literatures in the section.

2.1 Natural Image Dataset

The surface defect detection tasks are most related to the tasks of object detection in visual algorithms. Common benchmarks for visual object detection are constructed using natural images such as Pascal VOC [3] and MS-COCO [16]. These dataset are mostly balanced in terms of class distributions. The LVIS [8] dataset proposed a larger collection of images with a long-tailed distribution of classes. Our proposed dataset also has a long-tailed distribution with respect to classes. Unlike the general natural image datasets, our dataset also presents fine-grained class definition due to the nature of railway images.
Table 2: Categories statistics.

<table>
<thead>
<tr>
<th>Class</th>
<th>Running surface</th>
<th>Contact band</th>
<th>Dark Contact Band</th>
<th>Spalling</th>
<th>Crack</th>
<th>Corrugation</th>
<th>Grinding</th>
</tr>
</thead>
<tbody>
<tr>
<td># Boxes</td>
<td>1082</td>
<td>1093</td>
<td>113</td>
<td>12582</td>
<td>3785</td>
<td>3349</td>
<td>337</td>
</tr>
<tr>
<td># Images</td>
<td>1080</td>
<td>1087</td>
<td>769</td>
<td>1005</td>
<td>375</td>
<td>445</td>
<td>179</td>
</tr>
<tr>
<td># Large</td>
<td>1082</td>
<td>1092</td>
<td>773</td>
<td>1277</td>
<td>2965</td>
<td>3329</td>
<td>336</td>
</tr>
<tr>
<td># Medium</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>5147</td>
<td>784</td>
<td>17</td>
<td>1</td>
</tr>
<tr>
<td># Small</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>6148</td>
<td>36</td>
<td>3</td>
<td>0</td>
</tr>
<tr>
<td>Class</td>
<td>Fastening</td>
<td>Spike Screw</td>
<td>Set Screw</td>
<td>Indentation</td>
<td>Burning</td>
<td>Welded Joint</td>
<td></td>
</tr>
<tr>
<td># Boxes</td>
<td>757</td>
<td>502</td>
<td>414</td>
<td>307</td>
<td>41</td>
<td>14</td>
<td></td>
</tr>
<tr>
<td># Images</td>
<td>582</td>
<td>424</td>
<td>360</td>
<td>216</td>
<td>10</td>
<td>8</td>
<td></td>
</tr>
<tr>
<td># Large</td>
<td>750</td>
<td>475</td>
<td>400</td>
<td>4</td>
<td>41</td>
<td>14</td>
<td></td>
</tr>
<tr>
<td># Medium</td>
<td>7</td>
<td>27</td>
<td>14</td>
<td>237</td>
<td>0</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td># Small</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>66</td>
<td>0</td>
<td>0</td>
<td></td>
</tr>
</tbody>
</table>

2.2 Synthetic Corruption Dataset

There are also many datasets focusing on testing the robustness of deep-learning models under domain shift and image corruptions like ImageNet-C [10], CityScapes-C [18], and COCO-C [19]. However, the corruptions in these dataset are synthetic, generated using image processing techniques. Also, they are mainly used as the test set to test the robustness rather than the training set. In our dataset, the labeled dataset are well-curated, but the unlabeled set mat contains various real-world corruption, thus poses a new challenge for semi-supervised learning method.

2.3 Rail Defects Dataset

In the rail engineering domain, there are dataset focusing on the classification and detection of railway defects [23]. As for rail engineering, images are mostly in the form of atlas for manual reference. There are classification and detection [23] datasets of railway scene, as well as ultrasonic inspection datasets [11]. But still lacking of real-world datasets for rail surface defects. Faghih-Roohi etal. [4] collects and labels 100 x 50 resolution images in 6 defects classes. RSDDs datasets [6] contains 195 gray-scale images in 2 kinds of railway with segmentation mask. Feng etal. [5] collects thousands of images and annotate corrugation, fatigue and spalling in band region. Datasets above are all collected by high-speed linear scan cameras with low resolution and coarse-grained annotation. As a consequence, they all fail to drive the training of real-world robust deep learning algorithms.

Figure 1: Typical image capture and annotations.
3 The Rail-5k dataset

3.1 Rail Image Acquisition

The rail surface defects are mostly caused by the metal fatigue under the constant load from the wheel in high-speed section in a railway system. Rail images in the Rail-5k dataset were captured by specialized cameras mounted on inspection cars riding along the railway, making the lens 200 mm vertically away from the rail surface and focusing vertically downward. We exclude images with shadows or overexposure on the rail surface for railway experts to label. We collected annotations for 1100 RGB images with $3648 \times 2736$ pixels in resolution, covering scenarios as tunnel, elevated bridge, straight and curve line, inner and outer rail, before and after grinding or milling. Fig. 1 shows the map of a typical rail section that we collect images. Each dot represents an image.

We also collected 3k images from uncurated images of rail surfaces. These images contain unknown corruption and unseen objects in the labeled set. Fig. 3 shows some typical images in the unlabeled set.

In summary, our dataset contains two parts of data, the first part is the labeled subset with 1k labeled images, the second part is the unlabeled subset with 3k images. Thus our dataset can support both supervised and semi-supervised learning settings.

![Figure 2: Map of typical sample points.](image)

![Figure 3: First line is corruption images, second line is prediction results. It can be observed that there are many false positives.](image)

3.2 Fine-grained class definition and instance-level annotation

The annotations in our dataset were labeled by ten railway experts, each labeled images were at least checked by three experts. Based on the expert knowledge and railway standards, we use a fine-grained class definition and instance-level annotation paradigm for the railway defects detection. The labeling principle are listed in table 3.
Table 3: Annotation paradigm.

<table>
<thead>
<tr>
<th>Size</th>
<th>Boundary</th>
<th>Typical class</th>
<th>Annotation paradigm</th>
</tr>
</thead>
<tbody>
<tr>
<td>Large</td>
<td>clear</td>
<td>Rail surface, Fastener, Screw</td>
<td>external rectangular box(same as common detection)</td>
</tr>
<tr>
<td></td>
<td>obscure</td>
<td>Corrugation</td>
<td>wave valley of corrugation</td>
</tr>
<tr>
<td>Small</td>
<td>clear</td>
<td>Spalling, Indentation</td>
<td>stripped dent</td>
</tr>
<tr>
<td>Diffuse</td>
<td>sharp</td>
<td>Crack</td>
<td>union regions of small and dense boxes envelops cracking diffuse regions</td>
</tr>
</tbody>
</table>

Note that the crack area are sharp and thin objects with no clear edge boundary, we annotate with a segmentation mask.

3.3 Dataset Splitting

We randomly split 20% of the 1,100 labeled images in to the test set, the remaining images are used as the training set in the supervised setting. For the semi-supervised setting, we use the same test set to evaluate the performance for comparison.

4 Annotation Statistics

In this section, we present the statistics of our dataset. The statistics are presented in three aspects, namely the image and bounding box distribution among class, the bounding box sizes and aspect ratios, and the center point of annotated bounding boxes.

4.1 Class distribution

The number of images and annotations containing each classes. The Burning and welded joint are ignored in our experiments and benchmark because of their rare appearance. The imbalance ratio with respect to the number of bounding box between the most majority class and the most minority is 40.98, the imbalance ratio with respect to the number of images is 6.07.

4.2 Sizes and aspect ratios of bounding boxes

Bounding box annotations in our dataset vary dramatically in sizes and aspect ratios. There exist both tall and narrow objects as well as short and wide objects such as rail surface and contact band, normal square objects(fastener and screw). Besides, as shown in Figure 5, there are tremendous numbers of densely distributed small objects like spalling.

Figure 4: PR curve

Figure 5: Concrete and Constructions
4.3 Object positions

Figure 7 shows the distribution of objects’ center positions in our dataset. Because of the special
shooting paradigm, rail surfaces usually lie horizontally or vertically in images. As a consequence,
defects usually spread at the cross-zone in images.

![Figure 6: Width-height ratio of all annotations.](image)

![Figure 7: Center positions of all annotations.](image)

5 Pilot Study on the Rail-5k Dataset

In this section, we conducted comprehensive experiments in several aspects to investigate the
challenges and potential of the Rail-5k dataset. We trained an object detection model and a semantic
segmentation model on Rail-5k as our baselines and showed the challenging attributes of our dataset.
Additionally, We proposed a semi-supervised benchmark for object detection.

![Figure 8: Typical prediction results on testset.](image)

5.1 Benchmark for Detection

There are many popular detectors [20, 21, 15] on general object detection datasets. Recently,
many new methods have been proposed and achieved the state-of-the-art results in the MS-COCO
benchmark [16]. For example, YOLOv5 [13] is a light-weight model with mosaic augmentation
and Generalized Intersection over Union (GIoU) loss. In our experiments, we finetined Yolov5-s
as baseline on our dataset with MS-COCO pretraining. Detailed training settings are according to
data/hyp.finetune.yaml.

It can be noticed that the detector’s performance on crack is extremely low. This is because crack
is more a texture than an object without clear definition of separated instances. Thus, we chose to
tackle with this problem from another approach, which will be further discussed in Section 5.2.

---

2We implemented our experiments with Release v4.0 from https://github.com/ultralytics/yolov5/blob/develop/data/hyp.finetune.yaml
5.2 Benchmark for Crack

For cracking region, it is more of a texture and pattern than an object. Thus, we use segmentation to identify this class because the detection cannot recognize it well. We use Deeplabv3 [1] architecture with ResNet50 [9] backbone as segmentation model. The model is trained for 9000 iterations with IoU on background and 67.8% IoU on crack, which is much better than the detection performance. For cracking region, it is more of a texture and pattern than an object. Thus, we use segmentation to identify this class because the detection cannot recognize it well. We use Deeplabv3 [1] architecture with ResNet50 [9] backbone as segmentation model. The model is trained for 9000 iterations with a batch size of 16. We use SGD with momentum as the optimizer. Momentum and weight decay are set to 0.01, 1e-4 respectively. For the evaluation of our benchmark we choose the most common benchmark on segmentation, which is Intersection over Union(IoU). The model achieves 98.9% IoU on background and 67.8% IoU on crack, which is much better than the detection performance. DeepLabv3 can learn the main and obvious crack, but will ignore the tiny one.

<table>
<thead>
<tr>
<th>Class</th>
<th>Precision</th>
<th>Recall</th>
<th>AP@0.5</th>
<th>mAP@0.5:0.95</th>
<th>AP</th>
</tr>
</thead>
<tbody>
<tr>
<td>Rail Surface</td>
<td>77.5</td>
<td>99.1</td>
<td>98.9</td>
<td>90.6</td>
<td>96.6</td>
</tr>
<tr>
<td>Contact Band</td>
<td>60.2</td>
<td>97.7</td>
<td>94.5</td>
<td>71.9</td>
<td>96.3</td>
</tr>
<tr>
<td>Spalling</td>
<td>33.2</td>
<td>74</td>
<td>60</td>
<td>24.8</td>
<td>58.9</td>
</tr>
<tr>
<td>Corrugation</td>
<td>60.3</td>
<td>91.2</td>
<td>89.3</td>
<td>48.2</td>
<td>87.6</td>
</tr>
<tr>
<td>Grinding</td>
<td>21.4</td>
<td>38.8</td>
<td>24</td>
<td>7.4</td>
<td>24.1</td>
</tr>
<tr>
<td>Dark Contact Band</td>
<td>64.4</td>
<td>81.4</td>
<td>76.7</td>
<td>36.7</td>
<td>83.4</td>
</tr>
<tr>
<td>Fastener</td>
<td>47.5</td>
<td>91.3</td>
<td>83.8</td>
<td>62.9</td>
<td>86.1</td>
</tr>
<tr>
<td>Spike Screw</td>
<td>37.8</td>
<td>92.5</td>
<td>86.8</td>
<td>48.6</td>
<td>91.8</td>
</tr>
<tr>
<td>Set Screw</td>
<td>58.6</td>
<td>88.2</td>
<td>87.3</td>
<td>52.2</td>
<td>88.5</td>
</tr>
<tr>
<td>Indentation</td>
<td>0</td>
<td>0</td>
<td>0.7</td>
<td>0.1</td>
<td>16.2</td>
</tr>
<tr>
<td>Crack</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
</tbody>
</table>

Table 5: Metrics of baseline model for semi-supervised detection.

<table>
<thead>
<tr>
<th>Class</th>
<th>$s_{thr} = 0.6$</th>
<th>$s_{thr} = 0.7$</th>
<th>$s_{thr} = 0.8$</th>
<th>$s_{thr} = 0.9$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Rail Surface</td>
<td>98.1</td>
<td>98.0</td>
<td>98.1</td>
<td>97.7</td>
</tr>
<tr>
<td>Contact Band</td>
<td>78.4</td>
<td>77.9</td>
<td>77.1</td>
<td>77.0</td>
</tr>
<tr>
<td>Spalling</td>
<td>60.1</td>
<td>58.9</td>
<td>57.9</td>
<td>58.2</td>
</tr>
<tr>
<td>Corrugation</td>
<td>89.6</td>
<td>89.2</td>
<td>89.5</td>
<td>88.6</td>
</tr>
<tr>
<td>Grinding</td>
<td>23.0</td>
<td>23.6</td>
<td>23.5</td>
<td>22.1</td>
</tr>
<tr>
<td>Dark Contact Band</td>
<td>92.7</td>
<td>92.9</td>
<td>93.1</td>
<td>92.4</td>
</tr>
<tr>
<td>Fastener</td>
<td>86.5</td>
<td>86.1</td>
<td>85.8</td>
<td>83.2</td>
</tr>
<tr>
<td>Spike Screw</td>
<td>93.2</td>
<td>94.6</td>
<td>91.3</td>
<td>87.4</td>
</tr>
<tr>
<td>Set Screw</td>
<td>88.7</td>
<td>88.5</td>
<td>87.2</td>
<td>85.4</td>
</tr>
<tr>
<td>Indentation</td>
<td>15.9</td>
<td>16.4</td>
<td>13.4</td>
<td>15.3</td>
</tr>
<tr>
<td>Crack</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>mAP@0.5</td>
<td>63.29</td>
<td>63.27</td>
<td>62.43</td>
<td>61.55</td>
</tr>
</tbody>
</table>
5.3 Benchmark for Semi-supervised Learning

With additional 3k unlabeled images, we proposed a semi-supervised object detection benchmark. We present results in Table 5. These results are generated with simple pseudo label technique. We inferenced on the unlabeled images with YoloV5-s trained following strategy described in Section 5.1. Then we apply a confidence score threshold $s_{th}$ on all predictions and use remaining predictions as pseudo labels. Finally, we finetuned this model jointly on labeled images and unlabeled images with pseudo labels for 1 epoch and a base learning rate of 4e-4. Other training settings are the as ones in Section 5.1.

As shown Table 5, detectors usually perform worse after being finetuned under semi-supervision. This could be caused by corruption and noise in unlabeled images.

6 Conclusion

We introduce Rail-5k, a real-world dataset for rail surface defects detection. We capture rail images across China and provide fine-grained instance-level annotations. This dataset poses new challenges both in rail maintenance and computer vision. As a baseline, we provide a pilot study on Rail-5k using off-the-shelf detection models. In later versions, Rail-5k will include more images and patterns, as well as more defects categories and image modalities, such as 3D-scan or eddy current data. This would make Rail-5k an even more standardized and inclusive real-world dataset. We hope this dataset will encourage more work on improving visual recognition methods for rail maintenance, particularly on object detection and semantic segmentation for real-world, fine-grained, small, and dense defects.

References


Checklist

1. For all authors...
   (a) Do the main claims made in the abstract and introduction accurately reflect the paper’s contributions and scope? [Yes] And this work also fits to the scope of NeurIPS 2021 Datasets and Benchmarks Track.
   (b) Did you describe the limitations of your work? [Yes] This work only focus on rail surface defects, and the dataset indicates extreme label imbalance across categories along with real world corrupted images.
   (c) Did you discuss any potential negative societal impacts of your work? [Yes] This work helps to detect rail defects and save costs for maintenance. It will never do harm to society.
   (d) Have you read the ethics review guidelines and ensured that your paper conforms to them? [Yes] Our paper conforms all of the ethics rules.

2. If you are including theoretical results...
   (a) Did you state the full set of assumptions of all theoretical results? [N/A]
   (b) Did you include complete proofs of all theoretical results? [N/A]

3. If you ran experiments (e.g. for benchmarks)... 
   (a) Did you include the code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL)? [Yes] Codes have been uploaded on Github, see URL in appendix.
   (b) Did you specify all the training details (e.g., data splits, hyperparameters, how they were chosen)? [Yes] Data splits, baseline model hyperparameters, data augmentation, loss function... are all mentioned in paper and are reproducible.
   (c) Did you report error bars (e.g., with respect to the random seed after running experiments multiple times)? [Yes] No obvious bias were found through multiple baseline models and a series of ablation experiments.
   (d) Did you include the total amount of compute and the type of resources used (e.g., type of GPUs, internal cluster, or cloud provider)? [Yes] See Section 5.1

4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets...
   (a) If your work uses existing assets, did you cite the creators? [N/A]
   (b) Did you mention the license of the assets? [Yes] This dataset is licensed under CC BY-NC-ND 4.0 license.
   (c) Did you include any new assets either in the supplemental material or as a URL? [Yes] More data and annotations will be added in the URL.
   (d) Did you discuss whether and how consent was obtained from people whose data you’re using/curating? [Yes] We have reached an agreement with the authority that we could collect, process, and mining the data for academic purposes.
   (e) Did you discuss whether the data you are using/curating contains personally identifiable information or offensive content? [Yes] No personal identifiable information is contained and all geographic details have been erased.

5. If you used crowdsourcing or conducted research with human subjects...
   (a) Did you include the full text of instructions given to participants and screenshots, if applicable? [N/A]
   (b) Did you describe any potential participant risks, with links to Institutional Review Board (IRB) approvals, if applicable? [N/A]
   (c) Did you include the estimated hourly wage paid to participants and the total amount spent on participant compensation? [N/A]