Appendix

TLDR

Rail-5k is an object detection dataset with images and annotations in Pascal VOC format, with a mission to detect defects and accessories on the rail surface.

The dataset is licensed under CC BY-NC-ND 4.0.

You can access dataset or reproduce all results easily, and our team will keep maintenance in long-term.

Contents

1 URL of accessing Rail-5k dataset and code

URL of downloading All images and annotations.

URL-A: https://www.dropbox.com/sh/yzq1g3asjz9a1kt/AAC3yNBE4W11lSEgjw2vqfpta?dl=0

URL-B: https://drive.google.com/drive/folders/1iJmWtjx0i2l_iwX48C29e6-_0lnnbUUs?usp=sharing

URL of Zenodo documentation http://doi.org/10.5281/zenodo.4872619

URL of reproducing https://github.com/TommyZihao/Rail-5k-dataset

All results are easily reproducible.

Submitted to the 35th Conference on Neural Information Processing Systems (NeurIPS 2021) Track on Datasets and Benchmarks. Do not distribute.

2 How the data can be read

Rail-5k dataset is an object detection dataset in widely used Pascal VOC format.

Each labeled image has an corresponding .xml annotation file with the same file name.

You can visualize image and annotation with common software like labelimg or labelme.

3 Datasheet documentation(datasheet)

3.1 Motivation

For what purpose was the dataset created?

Rail surface is prone to rolling fatigue contact, crack, spalling, corrugation, and other defects under cyclic load from the wheel, which endangers comfort and safety. 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. Recently, computer vision makes it possible to recognize fine-grained damaged objects on rail surfaces, that is, to distinguish, classify and recognize each damage independently, to realize the millimeter-level measurement and meter-level localization.

Although vision patterns carry abundant quantitative information for rail maintenance, little progress has been made to recognize and evaluate rail defects due to the lack of high-quality image datasets. Meanwhile, problems in rail defects also bring new challenges in computer vision.

As a consequence, there is still a lack of high-quality image datasets that can drive the training of high-performance deep learning algorithms. This dataset is created to fill the need of detecting rail surface defects, to help rail maintenance, save costs, promote comfort and safety of railway transportation.

What task does this dataset solve?

Object detection and semantic segmentation for rail defects and accessories.

Who created the dataset?

Shanghai Key Laboratory of Rail Infrastructure Durability and System Safety, Tongji University, Shanghai 201804, China The Key Laboratory of Road and Traffic Engineering, Ministry of Education, Tongji University, Shanghai 201804, China

Who funded the creation of the dataset?

We would like to acknowledge the support of the National Natural Science Foundation of China (51678445,51878661) and the Fundamental Research Funds for the Central Universities.

3.2 Composition

What do the instances that comprise the dataset represent?

See Table 1.

How many instances are there in total?

5000+ images and 51 videos. 1100 images are annotated.

Does the dataset contain all possible instances or is it a sample (not necessarily random) of instances from a larger set?

The dataset samples typical defects over 10 years covering high-speed railway and subway across China. It is representative of different kinds of rail defects.

Table 1: Dataset categories statistics

Class	Running surface	Contact band	Dark Contact Band	Spalling	Crack	Corrugation	Grinding
# Boxes	1082	1093	773	12582	3785	3349	337
#Images	1080	1087	769	1005	375	445	179
# Large	1082	1092	773	1277	2965	3329	336
#Medium	0	0	0	5147	784	17	1
# Small	0	1	0	6148	36	3	0

Class	Fastening	Spike Screw	Set Screw	Indentation	Burning	Welded Joint
# Boxes	757	502	414	307	41	14
# Images	582	424	360	216	10	8
# Large	750	475	400	4	41	14
# Medium	7	27	14	237	0	0
# Small	0	0	0	66	0	0

What data does each instance consist of?

Class and Annotation boxes.

Is there a label or target associated with each instance?

Yes, 1100 images are annotated by ten rail experts through labelimg, and were checked by another two experts.

Based on expert knowledge and Chinese railway standards, a series of methods of fine-grained class definition and instance-level annotation for rail defects object detection are proposed.

Is any information missing from individual instances?

No.

Are relationships between individual instances made explicit (e.g., users' movie ratings, social network links)?

Yes. There are a lot of **class-translation** situations in the dataset.

For example, crack deteriorated into spalling, corrugation gradually form on the contact band.

Meanwhile, all defects are on the rail surface and contact band, all components and accessories(like fasteners and screws) are beyond the rail surface.

Are there recommended data splits (e.g., training, development/validation, testing)?

We randomly split 1,100 labeled images with 4/1 train/test ratio through stratified sampling. The Rail-5k dataset thus contains 877/223/3000+ train/test/unlabled images.

Are there any errors, sources of noise, or redundancies in the dataset?

It also contains real-world corrupted images with dark, overexposure, blur, other tools, different lens distance, category transition, different screws, which are infeasible for non-experts to annotate and recognize.

Is the dataset self-contained, or does it link to or otherwise rely on external resources (e.g., websites, tweets, other datasets)?

No.

Does the dataset contain data that might be considered confidential?

No.

Does the dataset contain data that, if viewed directly, might be offensive, insulting, threatening, or might otherwise cause anxiety?

No.

Does the dataset relate to people?

No.

3.3 Collection Process

How was the data associated with each instance acquired?

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. There should be no shadow or overexposure on the rail surface, and the angle of the auxiliary light is about 30°.See Figure 2.

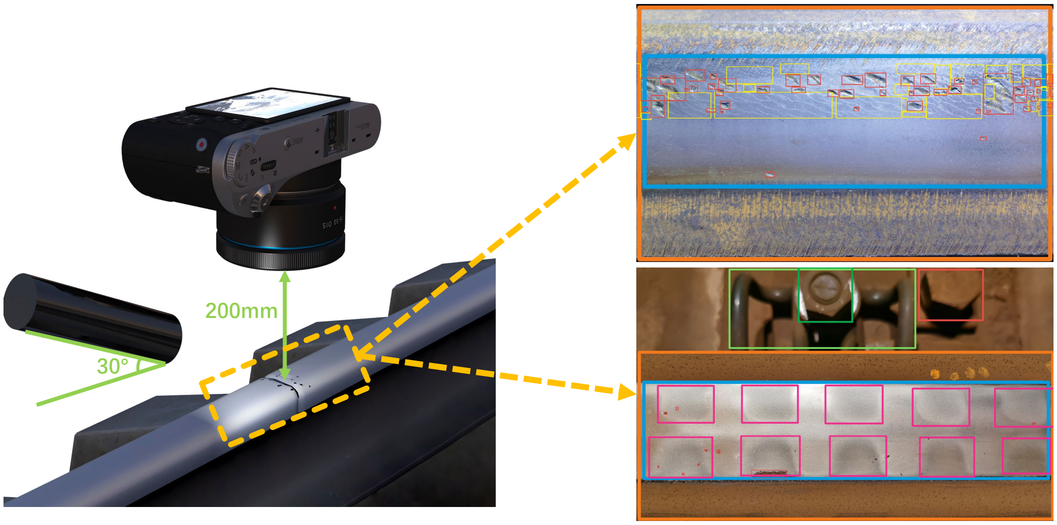


Figure 1: Image capture and annotation

The length or width direction of the image is parallel to the longitudinal edge of the rail surface, and the rail surface should occupy more than 60% of the image.

Through the above paradigm, 1100 RGB images (3648 x 2736 pixels) and 41 1080P high-resolution videos (1920 x 1080 pixels) were captured from sections with typical defects, covering high-speed railway and subway scenarios such as tunnel, elevated bridge, straight and curve line, inner and outer rail, before and after grinding or milling. See Table 3. The map shows a typical rail section that we collect images. Each dot represents an image. See Figure 2.

There are also 3000+ rail images but are not collected strictly with above paradigm. Some of them are corrupted.

So the dataset is divided into 2 parts: 1100 images with high quality and 3000+ images containing corruption. We annotate the first part and leave the second part as test images and semi-supervised assets.

What mechanisms or procedures were used to collect the data (e.g., hardware apparatus or sensor, manual human curation, software program, software API)?

SLR camera Mobile phone(iPhone XR) Inspection cars riding along the railway pushed by workers Manual checking.

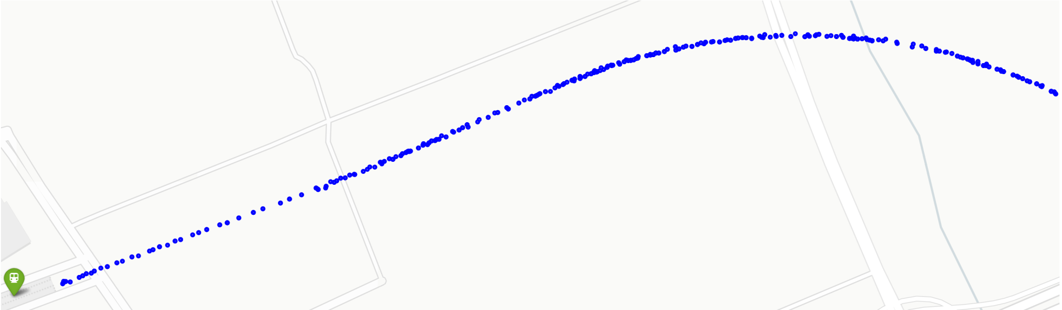
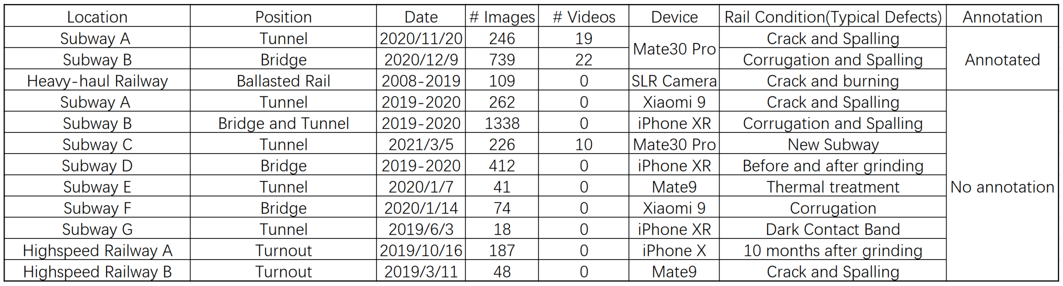


Figure 2: Sampling map



Location	Position	Date	# Images	# Videos	Device	Rail Condition(Typical Defects)	Annotation
Subway A	Tunnel	2020/11/20	246	19	Mate30 Pro	Crack and Spalling	Annotated
Subway B	Bridge	2020/12/9	739	22		Corrugation and Spalling	
Heavy-haul Railway	Ballasted Rail	2008-2019	109	0	SLR Camera	Crack and burning	
Subway A	Tunnel	2019-2020	262	0	Xiaomi 9	Crack and Spalling	No annotation
Subway B	Bridge and Tunnel	2019-2020	1338	0	iPhone XR	Corrugation and Spalling	
Subway C	Tunnel	2021/3/5	226	10	Mate30 Pro	New Subway	
Subway D	Bridge	2019-2020	412	0	iPhone XR	Before and after grinding	
Subway E	Tunnel	2020/1/7	41	0	Mate9	Thermal treatment	
Subway F	Bridge	2020/1/14	74	0	Xiaomi 9	Corrugation	
Subway G	Tunnel	2019/6/3	18	0	iPhone XR	Dark Contact Band	
Highspeed Railway A	Turnout	2019/10/16	187	0	iPhone X	10 months after grinding	
Highspeed Railway B	Turnout	2019/3/11	48	0	Mate9	Crack and Spalling	

Figure 3: Data acquisition

Who was involved in the data collection process (e.g., students, crowdworkers, contractors) and how were they compensated (e.g., how much were crowdworkers paid)?

Images and videos in Rail-5k dataset are all collected by students and teachers in Tongji University, most of the images are collected and annotated by myself(Zihao Zhang).

No compensate because we are self-driven volunteers to finish graduate project.

Were any ethical review processes conducted (e.g., by an institutional review board)?

Yes.

3.4 Annotation

Was any preprocessing/cleaning/labeling of the data done?

No preprocessing. No cleaning.

1100 images are annotated by ten rail experts through labelimg, and were checked by another two experts. Based on expert knowledge and Chinese railway standards, a series of methods of fine-grained class definition and instance-level annotation for rail defects object detection are proposed.

1-Large objects with clear boundary, such as rail surface and contact band. Label the external rectangular box as the annotation box, which is the same as common detection datasets.

2-Large objects with obscure boundary, such as corrugation. Corrugation usually presents periodic and lumpy changes on the rail surface. Label the wave valley as the annotation box.

3-Small objects with clear boundary, such as spalling. Label each stripped foreground area no matter how small it is.

4-Sharp and thin objects with no clear edge boundary, such as crack. Use a number of small and dense boxes which only contain crack as annotation, to envelope the whole cracking region. In other words, the union region of boxes is exactly the whole cracking region.

Is the software used to preprocess/clean/label the instances available?

Yes, labelimg is used to annotate object detection boxes.

3.5 Uses

Has the dataset been used for any tasks already?

No. This is a brand new dataset.

Is there a repository that links to any or all papers or systems that use the dataset?

Not yet.

What (other) tasks could the dataset be used for?

Measure and evaluate defects on the rail surface, such as:
Measure cracking area.
Measure the wavelength of corrugation.
Count spallings in different sizes and measure their areas.
Detect the defects and lost of fasteners and screws.

Is there anything about the composition of the dataset or the way it was collected and preprocessed/cleaned/labeled that might impact future uses?

No.

Are there tasks for which the dataset should not be used?

No. This work helps to detect rail defects and save costs for maintenance. It will never do harm to society.

3.6 Distrubution

Will the dataset be distributed to third parties outside of the entity (e.g., company, institution, organization) on behalf of which the dataset was created?

Yes, but with a series of strict terms of use.

How will the dataset will be distributed (e.g., tarball on website, API, GitHub)?

Through email application. We will review and decide whether to send datasets.

Does the dataset have a digital object identifier (DOI)?

Yes, http://doi.org/10.5281/zenodo.4872619

When will the dataset be distributed?

From June, 2021.

Will the dataset be distributed under a copyright or other intellectual property (IP) license, and/or under applicable terms of use (ToU)?

This dataset is licensed under CC BY-NC-ND 4.0 license.

Have any third parties imposed IP-based or other restrictions on the data associated with the instances?

No.

Do any export controls or other regulatory restrictions apply to the dataset or to individual instances?

Yes, through email application, we will review and decide whether to send datasets.

3.7 Maintenance

Who is supporting/hosting/maintaining the dataset?

Our team will support and maintain the dataset.

(Shanghai Key Laboratory of Rail Infrastructure Durability and System Safety, Tongji University)

How can the owner/curator/manager of the dataset be contacted (e.g., email address)?

407431120@qq.com, wqhuo2785@163.com

video channel: https://space.bilibili.com/1900783

Is there an erratum?

Yes, you can see all versions and update logs at http://doi.org/10.5281/zenodo.4872619

Will the dataset be updated (e.g., to correct labeling errors, add new instances, delete instances)?

Yes. Future versions of this dataset will include even more images, segmentation annotations as well as more channels.

Will older versions of the dataset continue to be supported/hosted/maintained?

Yes. We will give an erratum for old versions.

If others want to extend/augment/build on/contribute to the dataset, is there a mechanism for them to do so?

He can contact us by email and join our team.

He can both collect data with us and annotate images with us.

His work will be checked and reviewed by our expert team.

3.8 Impact and Challenges

This work helps to detect rail defects and save costs for maintenance. It will never do harm to society.

4 Author statement

We bear all responsibility in case of violation of rights, etc., and confirmation of the data license CC BY-NC-ND 4.0. This dataset is only used for non-profit moral purposes and academic research.

5 Hosting, licensing, and maintenance plan

Data application

This dataset is created to fill the need of detecting rail surface defects and accessories, to help rail maintenance, save costs, facilitate comfort and safety of railway transportation.

Data collection

We will get more images incorporating all kinds of railway scenarios including subway, high-speed railway, heavy-haul rail, trolley track, covering tunnel, bridge, ballasted and ballastless tracks, straight and curve lines, inner and outer tracks.

Data annotation

We will make more high-quality annotations in the future, including semantic segmentation of pix-level crack, fine-grained annotation of spike screws and set screws.

We will also propose methods of fine-grained class definition and instance-level annotation.

Data maintenance and distribution

Our team will organize a team consists of rail experts and students from Tongji University to keep the long-term preservation of the dataset.

Anyone who wants to download any version of Rail-5k may need to submit an application form through email.

Our team will check and verify each application.

Future version datasets, erratums, as well as papers will be available at http://doi.org/10.5281/zenodo.4872619

Contact Us

The Key Laboratory of Road and Traffic Engineering, Ministry of Education, Tongji University

Shanghai Key Laboratory of Rail Infrastructure Durability and System Safety, Tongji University

Email: 407431120@qq.com, wqhuo2785@163.com

Video channel: https://space.bilibili.com/1900783