

Faculty of Engineering and Information Technology  
University of Technology Sydney

# **Rail Infrastructure Defect Detection Through Video Analytics**

A thesis submitted in partial fulfillment of  
the requirements for the degree of  
**Doctor of Philosophy**

by

Huaxi Huang

February 2022

## CERTIFICATE OF ORIGINAL AUTHORSHIP

I, Huaxi Huang declare that this thesis, is submitted in fulfilment of the requirements for the award of Doctor of Philosophy, in the School of Electrical and Data Engineering, Faculty of Engineering and Information Technology at the University of Technology Sydney.

This thesis is wholly my own work unless otherwise referenced or acknowledged. In addition, I certify that all information sources and literature used are indicated in the thesis.

This document has not been submitted for qualifications at any other academic institution.

This research is supported by the Australian Government Research Training Program.

Production Note:

Signature: Signature removed prior to publication.

Date: **07/02/2022**

# Acknowledgments

First of all, I would like to express my most profound appreciation to my principal supervisor, Prof. Jian Zhang, for his professional guidance, reliable help, and perpetual support during my pursuit of the Ph.D. degree and three and a half years of research.

I would also like to express my sincere appreciation to my co-supervisor Prof. Qiang Wu and the collaborators: Dr. Junjie Zhang, Dr. Chang Xu, and Dr. Jingsong Xu, for not only their comments on revising my manuscript but also for the insightful guidance, which have incited me to improve my research ability and broaden my research horizon.

I am grateful to my colleagues and friends at the UTS Multimedia and Data Analytics Lab: Yazhou Yao, Xiaoshui Huang, Zhibin Li, Yongshun Gong, Lu Zhang, Anan Du, Lingxiang Yao, Guofeng Mei, Wenbo Xu, Litao Yu and all other labmates. I enjoyed the time we spent together.

Finally and most essentially, I would like to thank my parents. Their selfless support and continuing encouragement helped me defeat the obstacles I encountered during my Ph.D. study.

Huaxi Huang

February 2022 @ UTS

# Contents

<b>Certificate of Original Authorship</b> . . . . .	<b>i</b>
<b>Acknowledgment</b> . . . . .	<b>ii</b>
<b>List of Figures</b> . . . . .	<b>vi</b>
<b>List of Tables</b> . . . . .	<b>xi</b>
<b>List of Publications</b> . . . . .	<b>xiv</b>
<b>Abstract</b> . . . . .	<b>xv</b>
<b>Chapter 1 Introduction</b> . . . . .	<b>1</b>
1.1 Background . . . . .	1
1.2 Research Challenges . . . . .	4
1.2.1 Fine-grained Defect Recognition . . . . .	4
1.2.2 Few-shot Learning for Defect Recognition . . . . .	6
1.2.3 Summary . . . . .	8
1.3 Research Contributions . . . . .	8
1.4 Thesis Structure . . . . .	10
<b>Chapter 2 Literature Review</b> . . . . .	<b>12</b>
2.1 Automatic Industrial Defect Recognition . . . . .	12
2.2 Fine-grained Image Classification . . . . .	15
2.3 Generic Few-shot Learning . . . . .	17
2.4 Fine-grained Few-shot Learning . . . . .	20
2.5 Semi-Supervised Few-shot Learning . . . . .	22
<b>Chapter 3 RPSI Defect Recognition Using Fine-grained Deep Convolutional Neural Networks</b> . . . . .	<b>23</b>

3.1	Introduction . . . . .	23
3.2	RPSI Defect Dataset . . . . .	24
3.3	Methodology . . . . .	26
	3.3.1 STABLR Model . . . . .	26
	3.3.2 CNN Feature Acquisition . . . . .	29
3.4	Experiments . . . . .	31
	3.4.1 Experiment Setup . . . . .	31
	3.4.2 Experiments Results Analysis . . . . .	33
3.5	Summary . . . . .	35

## Chapter 4 Fine-grained Few-shot RPSI Defect Recognition

	<b>using Aligned Pairwise Bilinear Framework . . . . .</b>	<b>38</b>
4.1	Introduction . . . . .	38
4.2	Problem Definition . . . . .	44
4.3	Pairwise Alignment Bilinear Network . . . . .	45
	4.3.1 Pairwise Bilinear Pooling . . . . .	47
	4.3.2 Feature Alignment Loss . . . . .	48
4.4	Low-Rank Pairwise Alignment Bilinear Network . . . . .	49
	4.4.1 Low-Rank Pairwise Bilinear Pooling . . . . .	49
	4.4.2 Feature Alignment Layer . . . . .	52
	4.4.3 Comparator . . . . .	53
	4.4.4 Model Training . . . . .	54
	4.4.5 Network Architecture . . . . .	54
4.5	Target-Oriented Alignment Network . . . . .	56
	4.5.1 TOMM (Target-Oriented Matching Mechanism) . . . . .	60
	4.5.2 GPBP (Group pairwise Bilinear Pooling) . . . . .	61
	4.5.3 Comparator . . . . .	63
	4.5.4 Network Architecture . . . . .	63
4.6	Experiment . . . . .	65
	4.6.1 Datasets . . . . .	65
	4.6.2 Experimental Setup . . . . .	67

---

4.6.3	Experimental Results for PABN and LRPABN on Generic Fine-grained Datasets . . . . .	70
4.6.4	Experimental Results for PABN and LRPABN on the RPSI Defect Dataset . . . . .	72
4.6.5	Experimental Results for TOAN on RPSI Defects and Fine-grained Datasets . . . . .	77
4.6.6	Ablation Studies . . . . .	79
4.7	Summary . . . . .	93
<b>Chapter 5</b>	<b>Poisson Transfer Network for Semi-supervised Few-shot RPSI Defect Recognition . . . . .</b>	<b>94</b>
5.1	Introduction . . . . .	94
5.2	Methodology . . . . .	97
5.2.1	Problem Definition . . . . .	97
5.2.2	Representation Learning . . . . .	98
5.2.3	Poisson Label Inference . . . . .	100
5.2.4	Proposed Algorithm . . . . .	103
5.3	Experiments . . . . .	105
5.3.1	Datasets . . . . .	105
5.3.2	Implementation Details . . . . .	106
5.3.3	Experimental Results . . . . .	107
5.3.4	Ablation Study . . . . .	112
5.3.5	Inference Time . . . . .	114
5.4	Summary . . . . .	114
<b>Chapter 6</b>	<b>Conclusions and Future Work . . . . .</b>	<b>116</b>
6.1	Conclusions . . . . .	116
6.2	Future Work . . . . .	118
	<b>Bibliography . . . . .</b>	<b>119</b>

# List of Figures

1.1	An example of Railway Power Supply Infrastructure. We labeled two pieces of Dropper equipment with red and green boxes. The green one presents a standard Dropper equipment where the red one contains a defect Dropper equipment that is disconnected at the end of line 5 (L5). L1 (L2/3/4/5) indicates the power supply cable. . . . .	2
1.2	An example of Splice defect. (a) and (b) show V-wear defects in connection of wire and splice labeled by red boxes, (c) and (d) show standard Splices. (e) is the magnification of the first four images details. The first row of (e) shows the defect part of (a) which is a V-wear, the second row shows the defect part of (b) and the remaining rows are samples of standard parts of Splice (c)' bottom part and (d) (both top and bottom parts) which we draw with green lines. . . . .	5
3.1	An example of railway infrastructure defects categories. (a) and (b) represent Knuckle, (c) and (d) represent Kline, (e) and (f) represent Dropper1, and (g) as well as (h) represent Dropper2. The detailed parts to distinguish defects are labeled with red boxes. . . . .	26

3.2	The complete architecture of the STABLR model. STABLR can be divided into two parts: STN and Low-Rank BCNN. STN can learn the invariant representation of the dataset and Low-Rank BCNN can capture the fine-grained features of the input images. . . . .	27
3.3	P-R curves for all classes in STABLR models. . . . .	36
4.1	The high inter-class visual similarity and significant intra-class variations in FGFS tasks are more rigorous than general FG tasks. Some Herring gull and western gull images have similar visual appearances, which indicates the subtle inter-class variance. However, in each class, gulls present different postures with different backgrounds, which brings significant intra-class variance. . . . .	40
4.2	The proposed Aligned Pairwise Bilinear Framework (APBF) in the five-way-one-shot fine-grained image classification. The support set contains five labeled samples for each category (marked with numbers) and the query image labeled with a question mark. The APBF can be divided into four components: Encoder, Alignment Layer, Pairwise Bilinear Pooling, and Comparator. The Encoder extracts coarse features from raw images. Alignment Layer matches the pairs of support and query. Pairwise Bilinear Pooling acts as a fine-grained extractor that captures the subtle features. The Comparator generates the final results. . . . .	42
4.3	The pipeline of PABN under the one-shot fine-grained image recognition setting. There are three parts of PABN: Encoder, Fine-grained Features Extractor, and Comparator. Encoder extracts coarse features from raw images. Fine-grained Extractor captures the subtle features further. Comparator produces the final classification results. . . . .	46



4.4 Detailed network architectures used in LRPABN. (a) The Embedding network with Alignment Layer. (b) Low-Rank Pairwise Bilinear Pooling Layer. (c) The Comparator.  $I_i$  represents the query image, while  $I_j$  is the support image,  $x_i, x_j$  are the embedded image features and  $b_{i,j}$  represents the comparative bilinear feature.  $y_i$  is the predicted label by the comparator. 55

4.5 The overview of proposed TOAN in the N-way-one-shot FGFS task, other support samples are omitted (replaced by  $N$ ). The model consists of three parts: the feature embedding  $f_\theta$  learns the convolved features, the fine-grained relation extractor  $g_\omega$  generates bilinear features, and the comparator  $C_\phi$  maps the query to its ground-truth class.  $g_\omega$  contains target-orientated matching mechanism (TOMM) and group pairwise bilinear pooling (GPBP), TOMM aims at reformulating the features of support image to match the query image feature in the embedding space through the cross-correction attention mechanism, while GPBP is designed to extract discriminative second-order features by incorporating the channel grouping. With TOMM and GPBP,  $g_\omega$  learns to generate robust bilinear features from support-query pairs. PBP stands for the pairwise Bilinear Pooling, and we use different colors to indicate the feature maps in GPBP. . . . . 57

4.6 The architecture of fine-grained relation extractor, the left figure denotes TOMM, and the right one represents the GPBP operation.  $A$  and  $B$  indicate the embedded support sample and query sample respectively,  $Z$  is the fine-grained relation. . 64

4.7 Sample visual classification results of comparing methods over CUB dataset. All the approaches use the same data batch under the five-way-one-shot setting, and for each class, we randomly select five query images as the testing data. We adopt five colors to label the support classes separately. As to the query images, we label the images with the color corresponding to the class label predicted by different models. . . . . 75

4.8 The pairwise bilinear feature dimension selection experiment. In each sub-figure, the horizontal axis denotes the dimension of the pairwise bilinear feature and the vertical axis represents the test accuracy rate. 4.8(a) is the one-shot experiment and 4.8(b) is the five-shot experiment on CUB. . . . . 82

4.9 Visualization of the comparative feature generated by different fusion mechanism in 2D space using t-SNE. Each dot represents a query image that is numeric and marked with different colors according to the real labels. For each class, we randomly select thirty query images to conduct this experiment. The visualization is based on the CUB data set under the five-way-five-shot setting. (a) shows results conducted by RelationNet, (b) shows the result conducted by LRPABN<sub>cpt</sub>, and the dimension of the comparative bilinear feature is 128, denoted as LRPABN-Dim-128, (c) shows the result conducted by PABN+<sub>cpt</sub> model and (d) shows the result conducted by LRPABN<sub>cpt</sub>, and the dimension of the comparative bilinear feature is 512, denoted as LRPABN-Dim-512. . . . . 83

4.10 TOMM Visualization, the first image in each row (except for the first row) represents the support image, and the remaining images in the row are the aligned results of the support image, which are matched to each query image (in each column from the first row). . . . . 87

---

4.11	Ablation studies about the proposed GPBP, including semantic channel grouping validation 4.11(a) and feature dimension selection 4.11(b). For each group of validation experiments, we show the 1-shot and 5-shot results. . . . .	88
4.12	t-SNE visualization of the features learned by TOAN, five classes are randomly selected, and in each class, 30 query images are randomly chosen. Different colored numbers are used to denote different classes, <i>i.e.</i> , red zero represents the white-necked raven, blue one denotes the blacked capped vireo, green two represents the Laysan albatross, purple three denotes the nighthawk, and yellow four denotes the spotted catbird. Moreover, each sample is represented by its corresponding number. 4.9(a) and 4.9(b) show the t-SNE results of TOAN and RelationNet, respectively. . . . .	92
5.1	The overview of the proposed PTN. A feature embedding $f_{\theta_0}$ is pre-trained from the base-class set using standard cross-entropy loss first. This embedding is then fine-tuned with the external novel-class unlabeled data by adopting unsupervised transferring loss $\ell_{UT}$ to generate $f_{\theta}$ . Finally, we revise a graph model named PoissonMBO to conduct the query label inference. We also denote the Novel-Class Unlabel Set ( $U$ ), Support Set ( $S$ ), and Query Set ( $Q$ ) with different colors and shapes. . . . .	98
5.2	The five-way-one-shot and five-way-five-shot classification accuracy (%) using different number of extra unlabeled samples on the miniImageNet dataset. w/D means with distractor classes. . . . .	112

# List of Tables

3.1	Railway Power Supply Infrastructure (RPSI) Defects Dataset.	25
3.2	CNN features extractors for Railway Infrastructure Defects.	30
3.3	Detection Results for Railway Supply Power Infrastructure (RPSI) Defects Dataset.	34
4.1	The category partition for the four fine-grained datasets, which is the same as PCM	66
4.2	The class split of five datasets which is the same as	66
4.3	Few-shot classification accuracy (%) comparisons on four fine-grained data sets. The second-highest-accuracy methods are highlighted in blue color. The highest-accuracy methods are labeled with the red color. ‘-’ denotes not reported. All results are with 95% confidence intervals where reported.	71
4.4	Few-shot classification accuracy (%) comparisons on four fine-grained and RPSI Defect data sets. The highest-accuracy and second-highest-accuracy methods are highlighted in red and blue, respectively. All results are with 95% confidence intervals where reported.	73
4.5	Fine-grained Few-shot classification accuracy (%) comparisons on RPSI Defect and four FG benchmarks. All results are with 95% confidence intervals where reported. We highlight the best and second-best methods.	78

---

4.6	Ablation study of LRPABN with different components. The results are reported with 95% confidence intervals. Model size indicates the number of parameters for each model, the Test Time is the testing time for each input query image, and the Bilinear Dim represents the bilinear feature dimension of the each model. . . . .	80
4.7	Impact of input image size on FSFG. . . . .	81
4.8	The ablation study on TOMM and GPBP. The upper and lower parts of the table show the ablation study on TOMM and GPBP, separately. We incorporate each framework with TOMM and GPBP, we observe definite improvements (%). We also show the results of the whole model TOAN. . . . .	85
4.9	Ablation study of TOAN for other choices. Few-shot classification results (%) on four FG datasets. The lower parts of the table is the different backbone choices of TOAN. . . . .	86
4.10	Ablation study of TOAN for the output channel size of $d_\alpha, d_\beta$ . The table shows five-way few-shot recognition results (%) on the CUB dataset. . . . .	87
4.11	Investigation of model complexity. Model size indicates the number of parameters for each model, and the Test Time is the testing time for each input query image. . . . .	89
4.12	Investigation of model scalability. Model size indicates the number of parameters for each model. . . . .	90
5.1	The five-way-one-shot and five-way-five-shot image classification accuracy (%) on the RPSI Defect dataset with 95% confidence interval. . . . .	108
5.2	The five-way, one-shot and five-shot recognition accuracy (%) on the two datasets with 95% confidence interval. We mark the best performance in bold. The upper and lower parts of the table show the results on miniImageNet and tieredImageNet, respectively. . . . .	109

5.3	The five-way-one-shot and five-way-five-shot recognition accuracy (%) using various number of extra unlabeled samples on the miniImageNet dataset. PTN* denotes that we adopt PTN as the transductive model without fine-tune embedding. We mark the best results in bold. . . . .	110
5.4	Accuracy with various extra unlabeled samples for different semi-supervised few-shot methods on the <i>miniImageNet</i> dataset. All results are averaged over 600 episodes. We mark the best results in bold. . . . .	111
5.5	Distraction comparison on the <i>miniImageNet</i> dataset. . . . .	112
5.6	Distraction comparison on the <i>tieredImageNet</i> dataset. . . . .	113
5.7	Ablation studies about the proposed PTN, all methods are based on a pretrained embedding with 200 extra unlabeled samples each class on miniImageNet for five-way-one-shot and five-way-five-shot classification (%). Best results are in bold. . . . .	114
5.8	Mean inference time for the five-shot tasks on <i>miniImageNet</i> dataset. . . . .	115

# List of Publications

## Papers Published

- **Huaxi Huang**, Junjie Zhang, Jian Zhang, Qiang Wu, Chang Xu, *PTN: A Poisson Transfer Network for Semi-supervised Few-shot Learning*, in Proceeding of the 35th AAAI Conference on Artificial Intelligence (AAAI-21), pp: 1602-1609, 2021.
- **Huaxi Huang**, Junjie Zhang, Jian Zhang, Qiang Wu, Chang Xu, *TOAN: Target-Oriented Alignment Network for Fine-Grained Image Categorization with Few Labeled Samples*, IEEE Transactions on Circuits and System for Video Technology (TCSVT), vol: 32, pp: 853-866, 2022.
- **Huaxi Huang**, Junjie Zhang, Jian Zhang, Jingsong Xu, Qiang Wu. *Low-Rank Pairwise Alignment Bilinear Network For Few-Shot Fine-Grained Image Classification*. IEEE Transactions on Multimedia (TMM), vol: 23, pp: 1666-1680, 2021.
- **Huaxi Huang**, Junjie Zhang, Jian Zhang, Qiang Wu and Jingsong Xu, *Compare More Nuanced: Pairwise Alignment Bilinear Network for Few-Shot Fine-Grained Learning*, in IEEE International Conference on Multimedia and Expo (ICME-19), pp. 91-96, 2019.
- **Huaxi Huang**, Jingsong Xu, Jian Zhang, Qiang Wu, et al. *Railway Infrastructure Defects Recognition using Fine-grained Deep Convolutional Neural Network*. in IEEE International Conference on Digital Image Computing: Techniques and Application, pp: 1-8, 2018.

# Abstract

Compared with the traditional railway infrastructure maintenance process, which relies on manual inspection by professional maintenance engineers, inspection through automatic video analytics will significantly improve the working efficiency and eliminate the potential safety concern by reducing physical contact between maintenance engineers and infrastructure facilities. However, the defect does not always have a stable appearance and involves many uncertainties exposed in the clutter environments. On the other hand, various brands of the same devices are used widely on the railway, which shows diverse physical models. Therefore, it creates many challenges to the existing computer vision algorithms for defect detection. In this thesis, two key challenges are abstracted with regard to video/image analytics using computer vision techniques for railway infrastructure defect detection, resulting from the fine-grained defect recognition and the limited labeled learning (few-shot learning). This thesis summarizes the works that have been conducted on utilizing different methods to solve the two challenges.

The first challenge is fine-grained defect recognition. For railway infrastructure defect inspection, damaged or worn equipment defects are usually found in some small parts compared to the whole object. That is, the differences between the defective ones and standard ones are fine-grained. How to find these subtle defects is a fine-grained recognition problem. This thesis proposes a bilinear CNNs model to tackle the defect detection problem, which effectively captures the invariant representation of the dataset and learns high-order discriminative features for fine-grained defect recognition.



Another challenge is the limited labeled data (few-shot learning). In many scenarios, obtaining abundant labeled samples is laborious. For example, in industrial defect detection, most defects exist only in a few common categories, while most other categories only contain a small portion of defects. Moreover, annotating a large-scale dataset of railway infrastructure defects is labor-intensive, which requires high expertise in railway maintenance. Thus, how to obtain an effective model with sparse labeled samples remains an open problem. To address this issue, this thesis proposes a framework to simultaneously reduce the intra-class variance and enlarge the inter-class discrimination for both fine-grained defect recognition and general fine-grained recognition under the few-shot setting. Three models are designed according to this framework, and comprehensive experimental analyses are provided to validate the effectiveness of the models. This thesis further studies the few-shot learning problem by mining the unlabeled information to boost the few-shot learner for defect/general object recognition and proposes a Poisson Transfer Model to maximize the value of the extra unlabeled data through robust classifier construction and self-supervised representation learning.