

# Improved Fine-Grained Representation Learning with Data Transformation

by Lianbo Zhang

Thesis submitted in fulfilment of the requirements for the

*Degree of Doctor of Philosophy*

under the supervision of

by A/Prof. Wei Liu and Prof. Dacheng Tao

University of Technology Sydney

Faculty of Engineering and Information Technology

November 2021

To my beloved parents  
*Chuanyong Zhang and Jianzhen Li*

## **Certificate of Original Authorship**

I, Lianbo Zhang declare that this thesis, is submitted in fulfilment of the requirements for the award of Doctor of Philosophy in the School of Computer Science/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: 23/11/2021

## Publications During Enrolment

1. **Lianbo Zhang**, Shaoli Huang, Wei Liu, and Dacheng Tao. “*Learning a Mixture of Granularity-Specific Experts for Fine-Grained Categorization.*” In Proceedings of the IEEE/CVF International Conference on Computer Vision, 2019.
2. **Lianbo Zhang**, Shaoli Huang, and Wei Liu. “*Intra-Class Part Swapping for Fine-Grained Image Classification.*” Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, 2021.
3. **Lianbo Zhang**, Shaoli Huang, and Wei Liu. “*Enhancing Mixture-of-Experts by Leveraging Attention for Fine-Grained Recognition.*” IEEE Transactions on Multimedia, 2021.
4. **Lianbo Zhang**, Shaoli Huang, Xinchao Wang, Wei Liu, and Dacheng Tao “*Structure Aware Feature Generation for Zero-Shot Learning.*” arXiv preprint arXiv:2108.07032, 2021.

## **Acknowledgements**

Throughout the doctoral study in University of Technology Sydney, I have received a great deal of support and assistance.

I would first like to thank my supervisors, A/Prof. Wei Liu and Prof. Dacheng Tao, for their invaluable advice, continuous support, and patience during my PhD study. Their immense knowledge and plentiful experience have encouraged me in all the time of my academic research and daily life. Their expertise was invaluable in formulating the research questions and methodology. Their insightful feedback pushed me to sharpen my thinking and brought my work to a higher level.

Besides my supervisors, I would like to offer special thanks to my advisor, Dr. Shaoli Huang, for his technical support in my study. I am grateful for his great patience in answering my questions, as well as improving my academic writing. He taught me how to find interesting ideas, develop solid algorithms, and write academic papers.

In addition, I am deeply grateful to my parents, who gave me unreserved love and support. They are always there for me hearing my experience of study and life, sharing my joys and sorrows. Finally, I could not have completed this thesis without the support of my friends, Shanshan Zhao, Zeyu Feng, Yu Cao, Xinyuan Chen, Chaoyue Wang, Liu Liu, Qi Zheng, Youjian Zhang, Yaxin Shi and Ying Li, who provides stimulating discussion as well as distractions to rest my mind outside of my research.

## Abstract

Fine-grained recognition is challenging in computer vision and artificial intelligence. It aims to identify under subcategories of given images but suffers from small inter-class variance and large intra-class variance along with multiple object scales and complex background, leading to a more complex problem space. Recently, deep neural networks have extensively promoted the development of fine-grained recognition. However, the existing methods still suffer from several issues, including data limitation, model interpretation, and performance. In this thesis, we propose several data-transformation models to address these challenges.

First, we develop a unified framework (MGN-CNN) based on a mixture of experts to promote diversity among experts by combing a gradually-enhanced learning strategy and a KullbackLeibler divergence based constraint. The strategy learns new experts on the dataset with prior knowledge from former experts and adds them to the model sequentially. At the same time, the introduced constraint forces the experts to produce diverse prediction distributions. These drive the experts to learn the task from different aspects, making them specialized in various subspace problems.

Second, we propose Intra-class Part Swapping (InPS) that produces new data by performing attention-guided content swapping on input pairs from the same class. Compared with previous approaches, InPS avoids introducing noisy labels and ensures a likely holistic structure of objects in generated images. We demonstrate InPS outperforms the most recent augmentation approaches in both fine-grained recognition and weakly object localization.

Finally, we explore fine-grained zero-shot learning and introduce a novel structure-aware feature generation scheme, termed SA-GAN, to explicitly account for the topological structure in learning both the latent space and the generative networks. This topology-preserving mechanism enables our method to significantly enhance the generalization capability on unseen-classes and consequently improve the classification performance.

# Contents

<b>Dedication</b>	<b>i</b>
<b>Certificate of Original Authorship</b>	<b>ii</b>
<b>Publications During Enrolment</b>	<b>iii</b>
<b>Acknowledgements</b>	<b>iv</b>
<b>Abstract</b>	<b>v</b>
<b>List of Figures</b>	<b>vii</b>
<b>List of Tables</b>	<b>viii</b>
<b>1 Introduction</b>	<b>1</b>
1.1 Objectives and Motivation . . . . .	1
1.2 Problems and Challenges . . . . .	3
1.3 Contributions and Thesis Outline . . . . .	4
1.3.1 Contributions . . . . .	4
1.3.2 Outline . . . . .	5
<b>2 Literature Review</b>	<b>7</b>
2.1 Fine-Grained Recognition . . . . .	7
2.1.1 Classic Methods . . . . .	8
2.1.2 Covolutional Neural Networks . . . . .	12
2.1.3 Feature-Encoding Convolutional Neural Networks . . . . .	13
2.1.4 Part-Based Convolutional Neural Networks . . . . .	15
2.1.5 Efficient Learning . . . . .	21
2.2 Data Transformation . . . . .	22
2.2.1 Mixing-Based Strategy . . . . .	23
2.2.2 Feature Space Transformation . . . . .	27
2.2.3 Attention-Based Transformation . . . . .	28
2.2.4 GAN-based Augmentation . . . . .	30
<b>3 Mixture of Granularity-Specific Experts</b>	<b>33</b>
3.1 Introduction . . . . .	33
3.2 Related Works . . . . .	36
3.3 Method . . . . .	37
3.3.1 Experts for Fine-Grained Recognition . . . . .	38

---

3.3.2	KL-Divergence based Penalizing Term . . . . .	40
3.3.3	Mixture of Experts . . . . .	41
3.4	Experiments . . . . .	41
3.4.1	Implementation Details . . . . .	42
3.4.2	Experiments Results . . . . .	43
3.4.3	Ablation study . . . . .	46
3.5	Conclusion . . . . .	48
<b>4</b>	<b>Intra-Class Part Swapping</b>	<b>49</b>
4.1	Introduction . . . . .	49
4.2	Related Works . . . . .	50
4.3	Method . . . . .	52
4.3.1	Attention Priors . . . . .	53
4.3.2	Intra-class Part Swapping . . . . .	53
4.4	Experiments . . . . .	56
4.4.1	Dataset . . . . .	56
4.4.2	Implementation Details . . . . .	58
4.4.3	Intra-Class Attention Analysis . . . . .	58
4.4.4	Weakly Supervised Localization . . . . .	59
4.4.5	Fine-Grained Classification . . . . .	61
4.4.6	Ablation Study . . . . .	62
4.5	Conclusion . . . . .	63
<b>5</b>	<b>Structure-Aware Feature Generation</b>	<b>64</b>
5.1	Introduction . . . . .	64
5.2	Related Works . . . . .	66
5.3	Method . . . . .	68
5.3.1	Notation . . . . .	68
5.3.2	Structure-Preserving Mapping . . . . .	68
5.3.3	Structure-Aware Feature Generation . . . . .	70
5.4	Experiments . . . . .	71
5.4.1	Dataset . . . . .	72
5.4.2	Implementation Details . . . . .	72
5.4.3	Comparing with the State-of-the-Art . . . . .	73
5.4.4	(Generalized) Few-shot Learning . . . . .	73
5.4.5	Ablation Studies . . . . .	75
5.5	Conclusion . . . . .	78
<b>6</b>	<b>Conclusion</b>	<b>83</b>
6.1	Summary of Achievements . . . . .	83
6.2	Future Works . . . . .	84
	<b>Bibliography</b>	<b>86</b>



# List of Figures

1.1	Examples of animals/plants species with similar appearance. (a) is from [1], (b) is from [2], (c) is from [3], and (d) is from [4]	1
1.2	Examples of medical images [5].	2
1.3	Examples of fashion clothes. (a) is from [6]. (b) is from [7, 8].	3
2.1	Learning Part-based One-vs-One Features (POOF) for bird classification.	8
2.2	Examples of pose clustering and subcategory clustering for <b>Back</b> , marked with red dot. In (a), green dots are used to indicate the visible neighboring parts.	9
2.3	Image similarity measurement on the poselet set.	10
2.4	Overview of the Deformable Part Descriptors(DPDs).	10
2.5	Inference process by the alignment.	11
2.6	Overview of the Hierarchical Part Matching (HPM) framework.	12
2.7	Part-based R-CNNs for fine-grained image classification.	13
2.8	Bilinear models using second order statistics (a) as input, Architecture of (b) full bilinear model, (c) compact bilinear, (d) low-rank bilinear model.	13
2.9	Illustration of the higher-order integration framework.	14
2.10	Illustration of the Hierarchical Bilinear Pooling (HBP) framework.	15
2.11	Overview of the Deep LAC system.	16
2.12	Network architecture of Part-Stacked CNN model.	17
2.13	Overview of the deep filter responses picking framework.	18
2.14	The framework of multi-attention convolutional neural network (MA-CNN).	18
2.15	The framework of Discrimiative Filter Learning (DFL-CNN).	19
2.16	Overview of the two-level attention model.	20
2.17	Overview of the NTS-Net.	20
2.18	Overview of the weakly complementary part learning.	21
2.19	Overview of the Trilinear Attention Sampling Network (TASN).	22
2.20	Overview of the Trilinear Attention Sampling Network (TASN).	22
2.21	Overview of the Between-Class Learning.	24
2.22	Overview of the SamplePairing data augmentation.	24
2.23	Overview of the non-linear mixing.	25
2.24	Overview of the RICAP.	26
2.25	Overview of results of MixUp, CutOut and CutMix on ImageNet classification, ImageNet localization, and Pascal VOC07 detection. CutMix improve the performance on various tasks.	26
2.26	System architecture of the SA	27
2.27	A Squeeze-and-Excitation block.	28
2.28	The overview of CBAM	29

2.29	A spacetime non-local block. . . . .	29
2.30	Overview of attention-augmented convolution. . . . .	30
2.31	Overview of DCGAN generator. . . . .	30
2.32	Overview of Cycle-GAN. . . . .	31
2.33	Overview of conditional adversarial net. . . . .	32
3.1	Overview of our framework, which consists of several experts and a gating network. Each expert learns with prior knowledge from the previous expert. The gating network determines the contribution of each expert to the final predictions. . . . .	34
3.2	Network structure. The proposed MGE-CNN consists of several expert sub-networks, each of which contains a feature representation learning and attention region extraction component. The first component uses two different Conv blocks with different pooling methods on top of a shared Conv block to extract different types of feature and then concatenate them to form the overall representation. The second one is the gradient-based attention module, which is used to extract attention region and transform the training data into a new one for the following expert. . . . .	35
3.3	Attention module. We back-propagate gradients from ground-truth (predictions at test time) to obtain gradient of last convolutional layer. The gradient is then global average pooled and weighted summarized with feature maps along channel to get attention maps. The attention maps provide prior knowledge for latter expert. . . . .	36
3.4	Visualization of the selected results from CUB-200-2011 and Stanford Cars using proposed MGE-CNN. CAM is the class specific attention map. We remap each attention map back to match origin image. For each dataset, the first, third and fifth columns shows the input images to three experts, and the second, fourth and the last columns correspond to attention maps. . . . .	45
3.5	Visualization of the top-3 highest activation maps on selective exemplars from CUB200-2011 . . . . .	48
4.1	Comparison of Cutout, Mixup, CutMix, and the proposed method. Note that there is label mixing in MixUp and CutMix, and CutMix produces a new label based on category area. This might lead to noisy labels; for example, although Eared Grebe dominates ground-truth label, the output is visually more like California Gull to a human. In terms of object structure, Cutout and CutMix cause structure corruption; Mixup combines two input images unreasonably. Instead, our method generates more reasonable samples and clean supervision information. . . . .	49
4.2	Overview of our network architecture. InPS takes positive image pairs as input and then construct an attention pool using multiple-level features. After that, an attention pair is randomly selected before deploying a threshold to determine attended parts, which are swapped to generate synthetic images. . . . .	51
4.3	Weakly localization comparison under different threshold $\sigma$ on CUB-200-2011, Stanford-Cars and FGVC-Aircraft. . . . .	56

4.4	Qualitative comparison of the baseline (ResNet-50), Mixup, Cutout, Cut-Mix and InPS for weakly supervised object localization task on CUB-200-2011 dataset. Ground truth and predicted bounding boxes are denoted as green and red, respectively. . . . .	59
5.1	We quantitatively measure the average change of feature-prototype distance between the original visual space and the latent space on CUB dataset [4]. W-dist and L2-dist respectively denote Wasserstein distance and Euclidean distance. CLS denotes the classification loss, Center denotes the center constraint, and SP denotes the introduced structure-preserving constraint. A higher value indicates a greater change in geometric structure. . . . .	64
5.2	Comparison of different visual structure constraints for feature generation. (a) NE constraint [9] aims to maintain the neighbourhood structure between the visual and GAN space. (b) Center alignment [10] clusters fake samples to find visual centers and align the fake centers with the that of real ones. Here, $x^c$ and $\tilde{x}^c$ denote the class centers. (c) The proposed SA-GAN. Compared with existing methods, besides the difference in structure definition, our approach can better maintain the original structure information by using the mapped rather than the newly calculated prototype as a reference. Moreover, our method incorporates the prototype as condition input into the discriminator, which is more effective than adding a constraint loss to enforce the GAN to consider structure information. This is because the discriminator is usually the key to update the generator. . . . .	65
5.3	The framework of the proposed GZSL method. The latent feature $x_m$ is extracted by the mapping sub-network. The generator $G(\cdot)$ synthesizes new features $\tilde{x}_m$ based on class embedding $a$ and random noise $z$ . The discriminator $D(\cdot)$ tries to distinguish between real and fake instances by measuring the relationship with the class embedding $a$ and prototype $x_m^c$ . The second generator $G_2$ tries to recover the original visual features from the synthetic latent ones. . . . .	67
5.4	FSL and GFSL results on CUB dataset with increasing number of training samples per novel class. . . . .	74
5.5	(Generalized) Few-shot learning on the SUN dataset . . . . .	74
5.6	Impact of the number of synthetic instances on the CUB dataset. . . . .	76
5.7	Impact of the latent dimension in terms of ZSL, U, S, H on four datasets. . . . .	77
5.8	Comparing VAEGAN and SA-VAEGAN using t-SNE embeddings of the generated feature on SUN. The top row illustrates VAEGAN, and the bottom row shows our method. The symbol $\bullet$ denotes the instance of seen classes, and $\times$ denotes the instance of unseen classes. . . . .	78
5.9	Comparing VAEGAN and SA-VAEGAN using t-SNE embeddings of the generated feature on CUB. The top row illustrates VAEGAN, and the bottom row shows our method. The symbol $\bullet$ denotes the instance of seen classes, and $\times$ denotes the instance of unseen classes. . . . .	78
5.10	Model performances on the CUB dataset with different coefficients . . . . .	79

# List of Tables

3.1	Comparison of different methods on CUB-200-2011 (CUB) and Stanford-Cars (Car) with out extra annotations. . . . .	44
3.2	Comparison of different methods on Flowers-102 (Flower) and NABirds without extra annotations. . . . .	44
3.3	Comparing the effectiveness of KL-divergence constraints on CUB-200-2011. KL denotes expert with KL-divergence constraint. . . . .	46
3.4	Comparing the effectiveness of large and small part information on CUB-200-2011. . . . .	46
3.5	Experiments results using different threshold on CUB-200-2011. We only illustrate results combing two experts. . . . .	47
4.1	Dataset Statistics of CUB-200-2011, Stanford-Cars and FGVC-Aircraft. . . . .	56
4.2	Effectiveness of positive pair and attention pool on CUB-200-2011 . . . . .	56
4.3	Weakly supervised object localization comparison of state-of-the-art mixing-image approaches on CUB-200-2011, Stanford-Card, and FGVC-Aircraft. . . . .	57
4.4	Classification comparison of baseline(ResNet-50) and state-of-the-art augmentation methods (Mixup, Cutout, CutMix) on CUB-200-2011, Stanford-Cars, and FGVC-Aircraft. . . . .	57
4.5	Performance comparison with state-of-the-art methods on CUB200-2011, Stanford-Cars and FGVC-Aircraft. . . . .	60
4.6	Performance of middle-level representation on CUB200-2011, Stanford-Cars and FGVC-Aircraft. . . . .	61
4.7	Performance comparison in terms of classification accuracy (Acc) under different $\alpha, \beta$ on CUB-200-2011 dataset. . . . .	62
5.1	Statistics of five benchmark datasets used in the experiments, in terms of class embedding dimensions $K_a$ , number of seen classes $Y_s$ , number of unseen classes $Y_u$ , number of training samples $X^{tr}$ , numbers of test seen instances $X_s^{te}$ and unseen instances $X_u^{te}$ . . . . .	72
5.2	Comparing the proposed method with state-of-the-art methods on four benchmarks. We report average per-class top-1 accuracy for unseen (U) classes and seen (S) classes and their harmonic mean (H) in percentage. The best results are highlighted. . . . .	80
5.3	Classification accuracy (%) of conventional zero-shot learning for standard split (SS). The best results are highlighted. . . . .	81
5.4	Component contribution in terms of generalized zero-shot learning (H) on CUB dataset. SP-Map is structure-preserving mapping, mWGAN indicates structure-aware feature generation in the mapped space, and rWGAN reconstructs original visual features from latent ones. . . . .	81

---

5.5 Zero-shot learning using fine-tuned feature on CUB, AWA2 and SUN datasets. . . . .	82
--	----