

Data-efficient Visual Understanding via Deep Neural Networks

by Peike Li

Thesis submitted in fulfilment of the requirements for
the degree of

Doctor of Philosophy

under the supervision of Dr. Xin Yu

University of Technology Sydney
Faculty of Engineering and Information Technology

July 2022

CERTIFICATE OF ORIGINAL AUTHORSHIP

I, *Peike Li* declare that this thesis, is submitted in fulfilment of the requirements for the award of Doctor of Philosophy, in the *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.

SIGNATURE: Production Note:
Signature removed prior to publication. _____

DATE: 11th July, 2022

PLACE: Sydney, Australia

ACKNOWLEDGMENTS

Firstly, I would like to thank my principal supervisor Dr. Xin Yu, my co-supervisor Dr. Qian Peter Su, with my sincere and heartfelt gratitude and appreciation for the supervision journey that provided me with the guidance and counsel I needed to succeed in my Ph.D. program. They are great mentors in mapping my Ph.D. journey, advising on a research topic, and connecting me with the resources I need.

I would also like to thank my other mentors, collaborators and colleagues at University of Technology Sydney. I would like to thank Prof. Yi Yang, Prof. Yunchao Wei, Dr. Linchao Zhu, Dr. Ping Liu, Dr. Yu Wu, Dr. Jiaoxu Miao, Dr. Yawei Luo, Dr. Xiaohan Wang, Dr. Fan Ma, Dr. Xuanyi Dong, Dr. Yanbin Liu and many others. I was really fortunate to work with them and participate in intellectual conversations with them.

Lastly, I would like to thank my mother, Qiufen Li, my father, Hongru Li, and my wife, Qianyu Feng, for their support and love throughout my Ph.D. journey.

ABSTRACT

Despite the empirical and preliminary successes in computer vision, deep neural networks often require large-scale annotated training datasets. When applied to complex visual understanding problems in the real world, their performance is limited, since both data and annotations can be notoriously costly to collect, or may exist in various noisy or imperfect forms. Further, data annotating in such applications is also tedious to scale up, which demands highly skilled professionals, introducing challenges to use the cost-effective solutions, *e.g.*, crowdsourcing. Even worse, additional annotated data is always desired when the trained models need to be accordingly adapted to the dynamically changing environments. Thus, both the academic and industrial communities are calling for data-efficient deep learning algorithms.

In this thesis, we address the grand challenge of data-efficient and label-efficient visual understanding in realistic and imperfect real-world environments. To address this issue, we investigate deep learning approaches to leverage low-quantity training data and low-quality imperfect annotations. We propose a comprehensive suite of state-of-the-art approaches to tackle the data-efficient visual understanding from three directions, including : (1) applying *low-shot learning paradigms* that are intrinsically data-efficient, *e.g.*, few-shot learning or zero-shot learning. (2) exploiting imperfect labeled data to enable *learning with noise*. (3) *transferring prior knowledge* from the data-abundant domain into the data-hungry one. To demonstrate the effectiveness and efficiency in representative computer vision applications, extensive experiments are conducted on several dense prediction tasks, *e.g.*, human parsing, scene parsing, semantic segmentation, and face super-resolution.

LIST OF PUBLICATIONS

Related to the Thesis :

1. **P. Li**, Y. Xu, Y. Wei, and Y. Yang, “Self-Correction for Human Parsing,” in *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 2020.
2. **P. Li**, Y. Wei, and Y. Yang, “Consistent structural relation learning for zero-shot segmentation,” in *Advances in Neural Information Processing Systems (NeurIPS)*, 2020.
3. **P. Li**, Y. Wei, and Y. Yang, “Meta parsing networks: Towards generalized few-shot scene parsing with adaptive metric learning,” in *ACM International Conference on Multimedia (MM)*, 2020.
4. **P. Li**, X. Yu, and Y. Yang, “Super-resolving cross-domain face miniatures by peeking at one-shot exemplar,” in *IEEE International Conference on Computer Vision (ICCV)*, 2021.

Others :

5. **P. Li**, X. Dong, X. Yu, and Y. Yang, “When Humans Meet Machines: Towards Efficient Segmentation Networks,” in *The British Machine Vision Conference (BMVC)*, 2020.
6. **P. Li**, P. Pan, P. Liu, M. Xu, and Y. Yang, “Hierarchical Temporal Modeling with Mutual Distance Matching for Video Based Person Re-Identification,” in *IEEE Transactions on Circuits and Systems for Video Technology (TCSVT)*, 2020.
7. X. Pan, **P. Li**, Z. Yang, H. Zhou, C. Zhou, H. Yang, J. Zhou, and Y. Yang, “In-N-Out Generative Learning for Dense Unsupervised Video Segmentation,” in *ACM International Conference on Multimedia (MM)*, 2022.

-
8. Z. Yang, **P. Li**, Q. Feng, Y. Wei, and Y. Yang, “Going deeper into embedding learning for video object segmentation,” in *IEEE International Conference on Computer Vision Workshops*, 2019.
 9. Q. Feng, Z. Yang, **P. Li**, Y. Wei, and Y. Yang, “Dual embedding learning for video instance segmentation,” in *IEEE International Conference on Computer Vision Workshops*, 2019.
 10. X. Pan, H. Luo, W. Jiang, J. Zhang, J. Gu, and **P. Li**, “SFGN: Representing the Sequence with One Super Frame for Video Person Re-identification,” in *Knowledge-Based Systems*, 2022.

TABLE OF CONTENTS

List of Publications	vii
List of Figures	xiii
List of Tables	xvii
1 Introduction	1
1.1 Research Background	1
1.2 Data-efficient Visual Understanding	2
1.2.1 Low-shot Learning Paradigms for Data Efficiency	2
1.2.2 Data-efficient Learning from Imperfect Annotations	4
1.2.3 Data Efficiency via Prior Knowledge Transferring	4
1.3 Thesis Organization	6
2 Literature Review	9
2.1 Existing Data-hungry Visual Understanding Methods	9
2.2 Relevant Data-efficient Techniques	12
3 Self-correction for Human Parsing	15
3.1 Preface	15
3.2 Self-correction for Human Parsing	18
3.2.1 Overview	18
3.2.2 Single-person Human Parsing	19
3.2.3 Extension to Multi-person Human Parsing	22
3.3 Experiments: Single-person Human Parsing	23
3.3.1 Experiment Settings	24
3.3.2 Model-agnostic Study	25
3.3.3 Comparison with the State-of-the-art Approaches	27
3.3.4 Ablation Experiments	28

TABLE OF CONTENTS

3.3.5	Discussions	30
3.4	Experiments: Multiple-person and Video Human Parsing	32
3.4.1	Experiment Settings	32
3.4.2	Quantitative Results	34
3.4.3	Qualitative Results	35
3.5	Summary	36
4	Meta Parsing Networks: Towards Generalized Few-shot Scene Parsing with Adaptive Metric Learning	37
4.1	Preface	37
4.2	Task Definition	39
4.3	Meta Parsing Networks	41
4.3.1	Adaptive Deep Metric Learning	41
4.3.2	Contrastive Inter-class Distraction	44
4.3.3	Meta-training & Meta-testing	45
4.4	Experiment	46
4.4.1	Generalized Few-shot Scene Parsing Benchmarks	46
4.4.2	Generalization Ability of MPNet	48
4.4.3	Further Analysis of MPNet	50
4.5	Summary	52
5	Consistent Structural Relation Learning for Zero-Shot Segmentation	53
5.1	Preface	53
5.2	Preliminaries	54
5.3	Consistent Structural Relation Learning for Zero-Shot Segmentation . . .	56
5.3.1	Semantic-Visual Structural Generator	56
5.3.2	Consistent Structural Relation Learning	58
5.3.3	Training and Inference	59
5.4	Experiments	60
5.4.1	Experiment Settings	60
5.4.2	Comparisons with State-of-the-art Methods	61
5.4.3	Ablation Analysis	63
5.5	Summary	64
6	Super-Resolving Cross-Domain Face Miniatures by Peeking at One-Shot Exemplar	65

6.1	Preface	65
6.2	Task Definition: One-shot based FSR	67
6.3	Proposed Method	69
6.3.1	Domain Aware Pyramid-based FSR	70
6.3.2	Peeking at One-Shot Exemplar	73
6.3.3	Training and Inference	74
6.4	Experiments	75
6.4.1	Datasets and Evaluation Protocols	75
6.4.2	Implementation Details	75
6.4.3	Comparisons with the State-of-the-Art	76
6.4.4	Ablation Analysis	78
6.5	Summary	80
7	Conclusion	81
	Bibliography	83

LIST OF FIGURES

FIGURE	Page
3.1 Different types of label noises in ground-truth annotations. The upper row shows the original images. The lower row shows the original ground-truth labels. Different types of noisy labels are illustrated from left to right, (a) coarse annotation around the boundary area; (b) confused fine-grained categories, where the upper-cloth is mislabeled as the coat; (c) confused mirror categories, where the right leg is mislabeled as the left leg; (d) multiple-person occlusion. Annotation noises are marked in white dashed boxes.	16
3.2 Overview of the SChP pipeline. Starting from the warm-up initialization by training with inaccurate annotations, we design a cyclically learning scheduler to infer more reliable pseudo masks through iteratively aggregating the current learned model with the former optimal one in an online manner. Besides, those corrected labels can in turn to boost the model performance, simultaneously. In this way, the models and the masks get more robust and accurate during the self-correction cycles. Label noises are specially marked in white boxes.	18
3.3 The pipeline for the multiple-person human parsing and video human parsing task.	22
3.4 Model-agnostic study. The mIoU performance with different state-of-the-art models on LIP val set.	26
3.5 Visualization of SChP results on LIP val set. The first row shows the original input images. The middle row shows the ground-truth labels. Different human categories are shown in colors in the third row.	27

3.6	Examples from LIP train set during our self-correction process. Label noises like inaccurate boundary, confused fine-grained categories, confused mirror categories, multiple person occlusion are alleviated and resolved during the process. The boundaries of our corrected label are prone to be more smooth than the ground-truth label. Label noises are highlighted by white dotted boxes. Better zoom in to see the details.	28
3.7	Robustness of SChP against (a) different backbones and (b) context encoding modules. Experiments are conducted on LIP val set.	29
3.8	Performance curves <i>w.r.t</i> different training cycles. The mIoU, pixel accuracy and mean accuracy are depicted in the left, middle and right parts. All experiments are conducted on LIP val set.	31
3.9	Performance <i>w.r.t</i> different noise ratios on GTAV dataset.	32
3.10	Visualization results on MHP v2.0, CIHP and VIP val sets. All our results are depicted on the left part of each pair, while corresponding ground-truth labels are shown on the right side.	34
4.1	Compared to conventional few-shot segmentation task, the generalized few-shot scene parsing aims to segment complex scene scenario with multiple visual categories, where both seen and unseen categories are simultaneously considered. During meta-training stage, we first train the meta parser with the annotated images only on <i>seen</i> categories (e.g., road and vegetation), as indicated by the green colors. During meta-testing stage, given only one annotation image (one-shot) as guidance, the learned meta parser is then applied to segment both <i>seen</i> categories and <i>unseen</i> ones (e.g., car and person , as indicated by the blue colors.). Our target is to learn a meta parser that can generalize to both <i>seen</i> and <i>unseen</i> categories.	38
4.2	An overview of our MPNet. The ADML module aims to adaptively <i>learn</i> a transferable deep metric for dense comparison between support and query images. The CID module aims to encourage the feature discrepancy of different categories.	40
4.3	Illustration of the Adaptive Prototype Generation.	42
4.4	Dynamic instantiation for multiple categories	44
4.5	Performance <i>w.r.t</i> the number of support images (K-shot).	50
4.6	Visualization results of MPNet. Seen categories are indicated by the green colors while unseen categories by the blue colors.	51

4.7	Embedding visualization with t-SNE of (a) w/o CID (b) w/ CID. Different colors indicate different categories.	52
5.1	Illustration of CSRL. To achieve the goal of GZS3, we learn a generator to produce visual features from semantic word embeddings. Compared to (a) node-to-node generator, the proposed (b) structural generator explores the structural relations between seen and unseen categories to constrain the generation of unseen visual features.	55
5.2	The framework of the proposed CSRL. Our CSRL incorporates the feature generating and relation learning into a unified architecture. Given the semantic word embedding, CSRL generates visual features by alternately feature and relation aggregation. The proposed CSRL is trained under supervision from point-wise consistency on seen classes, pair-wise and list-wise consistency across seen and unseen classes.	57
5.3	Qualitative comparisons on Pascal-VOC dataset under the unseen-2 setting.	63
5.4	Relations between unseen (cow and motorbike) and seen categories.	63
6.1	Conventional FSR methods achieve good performance on the source dataset, but are prone to fail on the target dataset due to the domain gap. Our proposed method effectively adapts the model by leveraging only one-shot example. . .	66
6.2	Illustration of our DAP-FSR architecture. (a) The encoder network. Feature maps from different spatial resolution are up-sampled and concatenated as the multi-scale pyramid context. Each Adaptive Latent Encoding (ALE) module dynamically attends the multi-scale context to generate the latent representation \mathbf{w}_i . (b) The decoder network, where the HR images are generated based on the latent representations. (c) The Instance Spatial Transformer Network (ISTN) learns the style-invariant affine transformation matrix to adjust the unaligned LR images. (d) The detailed Adaptive Latent Encoding module, where the channel-wise feature attention is learned to adaptively capture the multi-scale information of the input images.	68
6.3	Compared to the style-transfer based method ASM [1] (left), given only one-shot target domain exemplar (ExtendedYaleB), our method (right) efficiently generates authentic target-style images from the source domain (CelebA). . .	72

LIST OF FIGURES

6.4	Comparisons with state-of-the-art methods on CelebA→ExtYaleB, CelebA→MultiPIE and MultiPIE→ExtYaleB benchmarks under the OSDA-FSR setting. Our method achieves high-quality, style-consistent HR faces and is also robust against unaligned LR inputs.	76
6.5	Comparisons with state-of-the-art methods on tiny faces in-the-wild [2] under real-world unconstrained conditions.	78
6.6	Comparisons with state-of-the-art methods on near-infrared (NIR) sensor captured faces [3].	79

LIST OF TABLES

TABLE	Page
3.1 Comparisons on the LIP validation set. The symbol † marks the single-scale testing result.	24
3.2 Comparisons on the PASCAL-Person-Part test set. The symbol † marks the single-scale testing result.	24
3.3 Comparison on the ATR test set. The symbol † marks the single-scale testing result.	25
3.4 The effect of our proposed model aggregation (MA) and label refinement (LR) strategy is evaluated on LIP val set.	30
3.5 The SCHP performance on Cityscapes & GTA5 Datasets.	30
3.6 Components analysis on val set of CIHP.	33
3.7 Comparison with state-of-the-arts on VIP val set. Our SCHP outperforms the other methods by a large margin. Specially, superior AP^r scores at high IoU thresholds are achieved by our method.	33
3.8 Comparison with state-of-the-arts on CIHP dataset.	35
3.9 Comparison with state-of-the-arts on MHP val set.	35
4.1 <i>GFSP-Cityscapes</i> benchmark splits. We only list the unseen categories, all rest are seen categories.	46
4.2 The comparison on <i>GFSP-Cityscapes</i> benchmark.	48
4.3 The comparison on <i>GFSP-Pascal-Context</i> benchmark.	49
4.4 The comparison of cross-domain experiments from <i>GFSP-Pascal-Context</i> to <i>GFSP-Cityscapes</i> benchmark.	49
4.5 Ablation analysis for the proposed modules of MPNet.	51
5.1 Generalized zero-shot semantic segmentation performance on Pascal-VOC dataset.	60
5.2 Generalized zero-shot semantic segmentation result on Pascal-Context dataset.	62

LIST OF TABLES

5.3	Ablation study of CSRL on Pascal-VOC.	63
6.1	Comparison with state-of-the-art methods. Results are reported on three benchmarks noted as source \rightarrow target. ‘Source only’ denotes the methods only using source dataset for training, while ‘one-shot’ denotes the methods exploring one-shot exemplar on the target dataset. \uparrow indicates that higher is better, and \downarrow that lower is better.	73
6.2	Ablations on different configurations of the network architecture (A,B,C) and different configurations of the adaptation algorithm (D,E,F). \uparrow indicates the higher the better, and \downarrow indicates the lower the better.	80
6.3	Comparisons on one-shot adaptation augmentation strategies. \uparrow indicates the higher the better, and \downarrow the lower the better.	80