UTS UNIVERSITY OF TECHNOLOGY SYDNEY

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 Porf. 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 thanks 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 Figures

Figure **Page**

List of Tables