UTS UNIVERSITY OF TECHNOLOGY SYDNEY

Generative and Discriminative Learning for Visual Matching

by Zhedong Zheng

Thesis submitted in fulfilment of the requirements for
the degree of

Doctor of Philosophy

under the supervision of Prof. Yi Yang and Dr. Liang Zheng

University of Technology Sydney
Faculty of Engineering and Information Technology

May 2021

CERTIFICATE OF ORIGINAL AUTHORSHIP

I, Zhedong Zheng, declare that this thesis, is submitted in fulfilment of the requirements for the award of Doctor of Philosophy, in the School of Computer Science at 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: 24, May, 2021

ABSTRACT

Generative and Discriminative Learning for Visual Matching

by

Zhedong Zheng

Visual matching aims to establish image correspondences across viewpoints. Given a query image, the visual matching system seeks to retrieve images containing the object of interest from non-overlapping viewpoints according to the similarity score. The visual matching task remains challenging because objects captured by different viewpoints often contain significant intra-class variations caused by background, viewpoint, object pose, etc. In this thesis, I present my research on combining generative learning with discriminative learning to build one robust visual matching system. First, due to lack of sufficient data to enhance robustness against input variations, generative learning is aimed at letting the model potentially "see" these variations (particularly intra-class variations) during training. With recent progress in the generative adversarial networks (GANs), generative models have become appealing choices to introduce additional augmented data for free. Second, discriminative learning is designed to formulate visual matching as a metric learning problem and adopt the discriminative optimization objective to learn the distance. With these objectives in mind, it motivates us to enable Convolutional Neural Network (CNN) to learn the mapping function to discriminate between different objects. In this thesis, I investigate two scientific problems of combining two learning strategies: 1) How to obtain high-quality generated data for subsequential training? 2) How to leverage the generated data to promote discriminative learning? To study the two problems, I explore improving learned visual representations by better leveraging the data from the following three aspects.

First, we present a semi-supervised pipeline that integrates GAN-generated images into discriminative learning. It contains a generative adversarial model for unsupervised data generation and a discriminative convolutional neural network for semi-supervised learning. Second, we observe that the generative pipelines are typically presented as standalone models, which are relatively separate from the discriminative learning models. To make the best of the two worlds, we further propose a learning framework that couples discriminative and generative learning. This design leads to a unified framework that enables the interactions between generative and discriminative modules in an end-to-end manner. Third, we further investigate different discriminative learning approaches on various data sources. Specifically, we study the feasibility of borrowing the knowledge from real-world vehicle images collected on the web and propose a two-stage learning strategy to minimize the domain gap between the web data and real-world data. Furthermore, we also explore the possibility of learning from synthetic data simulated by 3D engines. We propose a new geo-localization benchmark and build a strong and flexible baseline to learn from multi-view multi-source data.

In summary, this thesis studies and solves the critical challenges of data limitation and robust representation learning in visual matching. We show the benefits of leveraging the generative and discriminative learning in deep learning, which achieves better performance than previous methods.

Dissertation directed by Professor Yi Yang
The Australian Artificial Intelligence Institute (AAII), School of Computer Science

Acknowledgements

First, I would like to thank my supervisor, Prof. Yi Yang, for his encouragement, guidance, and patience. He has provided his perpetual supports to help me overcome difficulties and guided me to pursue challenging problems. He also taught me how to re-consider the same thing from different views, which is valuable to my future career. I have learned a lot under his caring supervision, and he has my deepest gratitude. Besides, I would like to thank my co-supervisor, Dr. Liang Zheng. He has provided me invaluable suggestions and consistent supports throughout my Ph.D. life. Thanks to Xiaodong Yang, Zhiding Yu, and Jan Kautz, my mentors when I interned at Nvidia Research, from whom I learned critical thinking and time management. Thanks to Mingyue Jiang and Xiao Tan, my mentors when I remotely interned at Baidu Research, from whom I learned the competition skills and knowledge. I also would like to thank all colleagues in the ReLER Lab at the University of Technology Sydney for their friendship and supports. I was fortunate to work with them.

Many thanks to Data to Decision CRC for supporting my research.

Finally, I would like to thank my parents Yanmin Yang and Gaoxin Zheng, my grandma Wenlan Tang for loves and supports over the years.

Zhedong Zheng
Sydney, Australia, 2021.

List of Publications

Journal Papers

J-1. **Zhedong Zheng**, Liang Zheng, Yi Yang (2018). A discriminatively learned cnn embedding for person reidentification. ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM), 14(1), 1-20. [**ESI Highly-cited Paper**]

J-2. **Zhedong Zheng**, Liang Zheng, Yi Yang (2018). Pedestrian alignment network for large-scale person re-identification. IEEE Transactions on Circuits and Systems for Video Technology (TCSVT), 29(10), 3037-3045. [**ESI Highly-cited Paper**]

J-3. **Zhedong Zheng**, Liang Zheng, Michael Garrett, Yi Yang, Mingliang Xu, Yi-Dong Shen (2020). Dual-Path Convolutional Image-Text Embeddings with Instance Loss. ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM), 16(2), 1-23.

J-4. **Zhedong Zheng**, Tao Ruan, Yunchao Wei, Yi Yang, Mei Tao (2020). VehicleNet: Learning Robust Visual Representation for Vehicle Re-identification. IEEE Transactions on Multimedia (TMM).

J-5. **Zhedong Zheng**, Yi Yang (2021). Rectifying Pseudo Label Learning via Uncertainty Estimation for Domain Adaptive Semantic Segmentation. International Journal of Computer Vision (IJCV), 1-15.

J-6. Zhun Zhong, Liang Zheng, **Zhedong Zheng**, Shaozi Li, Yi Yang (2018). Camstyle: A novel data augmentation method for person re-identification. IEEE Transactions on Image Processing (TIP), 28(3), 1176-1190. [**ESI Highly-cited Paper**]

J-7. Yan Huang, Jinsong Xu, Qiang Wu, **Zhedong Zheng**, Zhaoxiang Zhang, Jian

Zhang (2018). Multi-pseudo regularized label for generated data in person re-identification. IEEE Transactions on Image Processing (TIP), 28(3), 1391-1403.

J-8. Yutian Lin, Liang Zheng, **Zhedong Zheng**, Yu Wu, Zhilan Hu, Chenggang Yan, Yi Yang (2019). Improving person re-identification by attribute and identity learning. Pattern Recognition, 95, 151-161. [**ESI Highly-cited Paper**]

J-9. Qingji Guan, Yaping Huang, Zhun Zhong, **Zhedong Zheng**, Liang Zheng, Yi Yang (2020). Thorax disease classification with attention guided convolutional neural network. Pattern Recognition Letters, 131, 38-45.

J-10. Yutian Lin, **Zhedong Zheng**, Hong Zhang, Chenqiang Gao, Yi Yang (2020). Bayesian query expansion for multi-camera person re-identification. Pattern Recognition Letters, 130, 284-292.

J-11. Bingwen Hu, **Zhedong Zheng**, Ping Liu, Wankou Yang, Mingwu Ren (2020). Unsupervised eyeglasses removal in the wild. IEEE Transactions on Cybernetics, 1 - 13.

J-12. Tingyu Wang, **Zhedong Zheng**, Chenggang Yan, Jiyong Zhang, Yaoqi Sun, Bolun Zheng, Yi Yang (2021). Each Part Matters: Local Patterns Facilitate Cross-view Geo-localization. IEEE Transactions on Circuits and Systems for Video Technology (TCSVT).

Conference Papers

C-1. **Zhedong Zheng**, Liang Zheng, Yi Yang (2017). Unlabeled samples generated by gan improve the person re-identification baseline in vitro. In Proceedings of the IEEE International Conference on Computer Vision (ICCV) (pp. 3754-3762). [Spotlight]

C-2. **Zhedong Zheng**, Xiaodong Yang, Zhiding Yu, Liang Zheng, Yi Yang, Jan Kautz (2019). Joint discriminative and generative learning for person re-identification. In Proceedings of the IEEE conference on computer vision and pattern recognition (CVPR) (pp. 2138-2147). [Oral]

C-3. **Zhedong Zheng**, Yunchao Wei, Yi Yang (2020). University-1652: A Multi-view Multi-source Benchmark for Drone-based Geo-localization. In Proceedings of the ACM Multmieda (ACM MM) (pp. 1395–1403).

C-4. **Zhedong Zheng**, Yi Yang (2020). Unsupervised Scene Adaptation with Memory Regularization in vivo. In Proceedings of the International Joint Conference on Artificial Intelligence (IJCAI) (pp. 1076-1082).

C-5. Yawei Luo, **Zhedong Zheng**, Liang Zheng, Tao Guan, Junqing Yu, Yi Yang (2018). Macro-micro adversarial network for human parsing. In Proceedings of the European Conference on Computer Vision (ECCV) (pp. 418-434).

C-6. Zhun Zhong, Liang Zheng, **Zhedong Zheng**, Shaozi Li, Yi Yang (2018). Camera style adaptation for person re-identification. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (pp. 5157-5166).

C-7. Zhikun Huang, **Zhedong Zheng**, Chenggang Yan, Hongtao Xie, Yaoqi Sun, Jianzhong Wang, Jiyong Zhang (2020). Real-World Automatic Makeup via Identity Preservation Makeup Net. In Proceedings of the International Joint Conference on Artificial Intelligence (IJCAI) (pp. 652-658).

C-8. Chuchu Han, **Zhedong Zheng**, Changxin Gao, Nong Sang, Yi Yang (2021). Decoupled and Memory-Reinforced Networks: Towards Effective Feature Learning for One-Step Person Search. In Proceedings of the AAAI Conference on Artificial Intelligence (AAAI).

The source codes of most works can be found at `https://github.com/layumi` .

Contents

Certificate ii

Abstract iii

Acknowledgments v

List of Publications vi

List of Figures xiii

1 Introduction 1

1.1 Motivation 2

1.2 Approach 2

1.3 Contributions 3

1.4 Outline 5

2 Literature Review 7

2.1 Discriminative Learning 7

2.1.1 Person Re-identification 7

2.1.2 Vehicle Re-identification 8

2.1.3 Cross-view Geo-localization 9

2.2 Generative Learning 10

2.2.1 Generative Adversarial Networks 10

2.2.2 Dataset Augmentation 11

2.3 Related Works on Machine Learning 12

2.3.1 Semi-supervised Learning 12

2.3.2 Transfer Learning 13

3 Semi-supervised Learning with Generated Data 15

3.1 Introduction 15

3.2 Network Overview 18

3.2.1 Generative Adversarial Network 19

3.2.2 Convolutional Neural Network 20

3.3 The Proposed Regularization Method 21

3.3.1 Label Smoothing Regularization Revisit 21

3.3.2 Label Smoothing Regularization for Outliers 23

3.4 Experiment 25

3.4.1 Person Re-id Datasets 26

3.4.2 Implementation Details 27

3.4.3 Evaluation 28

3.4.4 Fine-grained Recognition 35

3.5 Summary 37

4 Multi-view Multi-source Image Matching 38

4.1 Introduction 38

4.2 Geo-localization Dataset Review 42

4.3 University-1652 Dataset 44

4.3.1 Dataset Description 44

4.3.2 Evaluation Protocol 46

4.4 Cross-view Image Matching 47

4.4.1 Visual Representations 47

4.4.2 Network Architecture and Loss Function 47

4.5 Experiment 50

4.5.1 Implementation Details 50

4.5.2 Geo-localization Results 50

4.5.3 Ablation Study and Further Discussion 55

4.6 Summary 60

5 Two-stage Progressive Learning 61

5.1 Introduction 61

5.2 Dataset Collection and Task Definition 64

5.2.1 Dataset Analysis 64

5.2.2 Task Definition 65

5.3 Methodology 67

5.3.1 Model Structure 67

5.3.2 Two-stage Progressive Learning 69

5.3.3 Post-processing 72

5.4 Experiment 75

5.4.1 Implementation Details 75

5.4.2 Quantitative Results 77

5.4.3 Further Evaluations and Discussion 79

5.5 Summary 85

6 Joint Discriminative and Generative Learning 86

6.1 Introduction 86

6.2 Methodology 90

6.2.1 Generative Module 91

6.2.2 Discriminative Module 94

6.2.3 Optimization 96

6.3 Experiment 97

6.3.1 Implementation Details 98

6.3.2 Generative Evaluations 99

6.3.3 Discriminative Evaluations 102

6.4 Summary 104

7 Conclusions and Future Work 107

7.1 Summary of Contributions 107

7.2 Future Directions 108

Bibliography 111

List of Figures

3.1 The pipeline of the proposed method. There are two components: a generative adversarial model [121] for unsupervised learning and a convolutional neural network for semi-supervised learning. “Real Data” represents the labeled data in the given training set; “Training data” includes both the “Real Data” and the generated unlabeled data. Our target is to learn more discriminative embeddings with the “Training data”. 16

3.2 The image distribution per class in the dataset Market-1501 [206], CUHK03 [76] and DukeMTMC-reID (Duke) [126, 217]. We observe that all these datasets suffer from the limited images per class. Note that there are only a few classes with more than 20 images. 19

3.3 Examples of GAN images and real images. (a) The top two rows show the pedestrian samples generated by DCGAN [121] trained on the Market-1501 training set [206]. (b) The bottom row shows the real samples in training set. Although the generated images in (a) can be easily recognized as fake images by a human, they still serve as an effective regularizer in our experiment. 20

3.4 The label distributions of a real image and a GAN-generated image in our system. We use a classical label distribution (Eq. 3.2) for the real image (left). For the generated image (right), we employ the proposed LSRO label distribution (Eq. 3.6), *e.g.*a uniform distribution on every training class because the generated image is assumed to belong to none of the training classes. We employ a cross-entropy loss that combines the two types of label distributions as the optimization objective (Eq. 3.7). 23

3.5 The newly generated images from a DCGAN model trained on DukeMTMC-reID (Duke) and CUB-200-2011. Through LSRO, they are added to the training sets of DukeMTMC-reID and CUB-200-2011 to regularize the CNN model. 28

3.6 Sample retrieval results on DukeMTMC-reID using the proposed method. The images in the first column are the query images. The retrieved images are sorted according to the similarity scores from left to right. The correct matches are in the blue rectangles, and the false matching images are in the red rectangles. DukeMTMC-reID is challenging because it contains pedestrians with occlusions and similar appearance. 34

4.1 It is challenging, even for a human, to associate (a) ground-view images with (b) satellite-view images. In this chapter, we introduce a new dataset based on the third platform, *i.e.*, drone, to provide real-life viewpoints and intend to bridge the visual gap against views. (c) Here we show two real drone-view images collected from public drone flights on Youtube [3, 31]. (d) In practice, we use the synthetic drone-view camera to simulate the real drone flight. It is based on two concerns. First, the collection expense of real drone flight is unaffordable. Second, the synthetic camera has a unique advantage in the manipulative viewpoint. Specifically, the 3D engine in Google Earth is utilized to simulate different viewpoints in the real drone camera. 39

4.2 (a) The drone flight curve toward the target building. When flying around the building, the synthetic drone-view camera could capture rich information of the target, including scale and viewpoint variants. (b) The ground-view images are collected from street-view cameras to obtain different facets of the building as well. It simulates real-world photos when people walk around the building. . 40

4.3 The basic model architectures for cross-view matching. Since the low-level patterns of different data are different, we apply multi-branch CNN to extract high-level features and then build the relation on the high-level features. (I) Model-I is a two-branch CNN model, which only considers the satellite-view and ground-view image matching; (II) Model-II is a two-branch CNN model, which only considers the satellite-view and drone-view image matching; (III) Model-III is a three-branch CNN model, which fully utilizes the annotated data, and considers the images of all three platforms. There are no "standard" methods to build the relationship between the data of multiple sources. Our baseline model applies the instance loss [215] and we also could adopt other loss terms, *e.g.*, triplet loss [14, 24] and contrastive loss [82, 175, 216]. 48

4.4 The test accuracy curves when using n training drone-view images per class, $n \in \{1, 3, 9, 27, 54\}$. The two sub-figures are the Rank@1 (%) and AP (%) accuracy curves, respectively. The orange curves are for the drone navigation (Satellite → Drone), and the blue curves are for the drone-view target localization (Drone → Satellite). 53

4.5 Qualitative image retrieval results. We show the top-3 retrieval results of drone-view target localization (left) and drone navigation (right). The results are sorted from left to right according to their confidence scores. The images in yellow boxes are the true matches, and the images in the blur boxes are the false matches. (Best viewed when zoomed in.) 54

4.6 Qualitative image search results using real drone-view query. We evaluate the baseline model on an unseen university. There are two results: (I) In the middle column, we use the real drone-view query to search similar synthetic drone-view images. The result suggests that the synthetic data in University-1652 is close to the real drone-view images; (II) In the right column, we show the retrieval results on satellite-view images. It verifies that the baseline model trained on University-1652 has good generalization ability and works well on the real-world query. 55

5.1 The motivation of our vehicle re-identification method by leveraging public datasets. The common knowledge of discriminating different vehicles could be transferred to the final model. 64

5.2 (a) The image distribution per class in the vehicle re-id datasets, *e.g.*, CityFlow [148], VehicleID [86] , CompCar [185] and VeRi-776 [95]. We observe that the two largest datasets, *i.e.*, VehicleID and CompCars, suffer from the limited images per class. (b) Here we also provide the image samples of the four datasets. The four datasets contain different visual biases, such as illumination conditions, collection places and viewpoints. 67

5.3 Illustration of the model structure. We remove the original classifier of the ImageNet pre-trained model, add a new classifier and replace the average pooling with the adaptive average pooling layer. The adaptive average pooling is to squeeze the output to the pre-defined shape (*i.e.*, 1×1). 68

5.4 Geometric Interpretation. Here we give a three-class sample to show our intuition. W_i denotes the class weight of the final linear classifier. In this example, the third class denotes one auxiliary class, which belongs to VehicleNet but the target domain. Therefore, in the Stage-II fine-tuning, we remove the auxiliary classes, including W_3. The cross-entropy loss of Stage-I pulls the samples with the same label together (close to either the relative weight W_1, W_2 or W_3). In this way, the positive pair is closer than the negative pair, while the samples are far from the decision boundary. Stage I, therefore, leads to a decent weight initialization to be used in Stage II with a large margin from decision boundary, when we leave out the auxiliary class, *i.e.*, the third class with W_3, from VehicleNet. . . 71

5.5 The inference pipeline for AICity Challenge Competition. Given one input image and the corresponding cropped image via MaskRCNN [40], we extract features from the trained models, *i.e.*, 8×SE-ResNeXt101 [48]. We normalize and concatenate the features. Meanwhile, we extract the camera prediction from the camera-aware model, *i.e.*, the fine-tuned DenseNet121 [51]. Then query expansion and camera verification are applied. Finally, we utilize the re-ranking technique [220] to retrieve more positive samples. 72

5.6 Qualitative image search results using the vehicle query images from the CityFlow dataset. We select the four query images from different viewpoints. The results are sorted from left to right according to the similarity score. The true-matches are in green, when the false-matches are in red. 81

5.7 Visualization of the activation heatmap in the learned model on VehicleNet. The vehicle images in every subfigure (a)-(c) are from the same vehicle ID. Noted that there do exist strong response values at the regions containing discriminative details, such as headlights and tire types. 81

5.8 The training losses of the two stages. Due to the large-scale data and classes, the first stage (left) takes more epochs to converge. Attribute to the trained weight of the first stage, the second stage (right) converge early. 85

6.1 Examples of generated images on Market-1501 by switching appearance or structure codes. Each row and column corresponds to different appearance and structure. 87

6.2 A schematic overview of DG-Net. (a) Our discriminative re-id learning module is embedded in the generative module by sharing appearance encoder E_a. A dash black line denotes the input image to structure encoder E_s is converted to gray. The red line indicates the generated images are online fed back to E_a. Two objectives are enforced in the generative module: (b) self-identity generation by the same input identity and (c) cross-identity generation by different input identities. (d) To better leverage generated data, the re-id learning involves primary feature learning and fine-grained feature mining. 91

6.3 Comparison of the generated and real images on Market-1501 across the different methods including LSGAN [104], PG^2-GAN [102], FD-GAN [33], PN-GAN [118], and our approach. This figure is best viewed when zoom in. Please attention to both foreground and background of the images. 97

6.4 Comparison of the generated images by our full model, removing online feeding (w/o feed), and further removing identity supervision (w/o id). 98

6.5 Example of image generation by linear interpolation between two appearance codes. 99

6.6 Examples of our generated images by swapping appearance or structure codes on the three datasets. All images are sampled from the test sets. 100

6.7 Comparison of success and failure cases in our image generation. In the failure case, the logo on t-shirt of the original image is missed in the synthetic image. 102

6.8 Analysis of the re-id learning related hyper-parameters α and β to balance primary and fine-grained features in training (left) and testing (right). 103