Towards robust sensing and recognition : from statistical learning to transfer learning

Yang, Jianfei

2020

Yang, J. (2020). Towards robust sensing and recognition: from statistical learning to transfer learning. Doctoral thesis, Nanyang Technological University, Singapore.

<https://hdl.handle.net/10356/145444>

<https://doi.org/10.32657/10356/145444>

This work is licensed under a Creative Commons Attribution-NonCommercial 4.0 International License (CC BY-NC 4.0).

Downloaded on 26 Apr 2023 22:49:35 SGT

Towards Robust Sensing and Recognition: From Statistical Learning to Transfer Learning

Jianfei Yang

School of Electrical & Electronic Engineering

A thesis submitted to the Nanyang Technological University
in partial fulfillment of the requirements for the degree of
Doctor of Philosophy

2020

Statement of Originality

I hereby certify that the work embodied in this thesis is the result of original research, is free of plagiarised materials, and has not been submitted for a higher degree to any other University or Institution.

July 16th, 2020

.....

Date

.....

Jianfei Yang

Supervisor Declaration Statement

I have reviewed the content and presentation style of this thesis and declare it is free of plagiarism and of sufficient grammatical clarity to be examined. To the best of my knowledge, the research and writing are those of the candidate except as acknowledged in the Author Attribution Statement. I confirm that the investigations were conducted in accord with the ethics policies and integrity standards of Nanyang Technological University and that the research data are presented honestly and without prejudice.

July 16th, 2020

.....

Date

.....

Prof. Xie Lihua

Authorship Attribution Statement

This thesis contains materials from 7 papers published or under review in the peer-reviewed journals and conferences in which I am listed as a co-author.

Chapter 3 contains research results published in

Jianfei Yang, Han Zou, Hao Jiang, and Lihua Xie. “Device-free occupant activity sensing using WiFi-enabled IoT devices for smart homes.” *IEEE Internet of Things Journal*, Vol. 5, no. 5 (2018): 3991-4002.

Jianfei Yang, Han Zou, Hao Jiang, and Lihua Xie. “CareFi: Sedentary behavior monitoring system via commodity WiFi infrastructures.” *IEEE Transactions on Vehicular Technology*, Vol. 67, no. 8 (2018): 7620-7629.

The contributions of the co-authors are as follows:

- I proposed the idea, designed the experiments and prepared the manuscript.
- Han Zou, Hao Jiang and I conducted the experiments.
- The manuscript was revised by Han Zou, Hao Jiang and Lihua Xie.

Chapter 4 contains results published in

Jianfei Yang, Han Zou, Yuxun Zhou, and Lihua Xie. “Learning Gestures From WiFi: A Siamese Recurrent Convolutional Architecture.” *IEEE Internet of Things Journal*, Vol. 6, no. 6 (2019): 10763-10772.

The contributions of the co-authors are as follows:

- I proposed the idea, designed the experiments and prepared the manuscript.
- Han Zou and I conducted the experiments.
- The manuscript was revised by Han Zou, Yuxun Zhou and Lihua Xie.

Chapter 5 has been accepted for publication as

Jianfei Yang, Han Zou, Shuxin Cao, Zhenghua Chen, and Lihua Xie. “MobileDA: Towards Edge Domain Adaptation.” *IEEE Internet of Things Journal* (2020).

The contributions of the co-authors are as follows:

- I proposed the idea, designed the experiments and prepared the manuscript.
- Han Zou and I conducted the experiments.
- The manuscript was revised by Han Zou, Shuxin Cao, Zhenghua Chen and Lihua Xie.

Chapter 6 contains a paper under review: [Jianfei Yang, Han Zou, Yuxun Zhou, and Lihua Xie](#). “Towards stable and comprehensive domain alignment: max-margin domain-adversarial training.”

The contributions of the co-authors are as follows:

- I proposed the idea, designed the experiments and prepared the manuscript.
- The experiments were conducted by me.
- The manuscript was revised by Han Zou, Yuxun Zhou and Lihua Xie.

Chapter 7 contains a paper that has been accepted for publication as

[Jianfei Yang, Han Zou, Yuxun Zhou, and Lihua Xie](#). “Asymmetric adversarial domain adaptation.” European Conference on Computer Vision (ECCV), 2020

The contributions of the co-authors are as follows:

- I proposed the idea, designed the experiments and prepared the manuscript.
- Han Zou, Yuxun Zhou and I discussed the idea.
- The experiments were conducted by me.
- The manuscript was revised by Han Zou, Yuxun Zhou and Lihua Xie.

Chapter 8 is a paper under review: [Jianfei Yang, Jiangang Yang, Han Zou, Shuxin Cao, Shizheng Wang, and Lihua Xie](#). “Cluster-level discrepancy minimization for imbalanced domain adaptation.”

The contributions of the co-authors are as follows:

- I proposed the idea, designed the experiments and prepared the manuscript.
- Jiangang Yang, Shuxin Cao and I conducted the experiments.
- The manuscript was revised by Han Zou, Shizheng Wang and Lihua Xie.

July 16th, 2020

.....

Date

.....

Jianfei Yang

Acknowledgements

First and foremost, I wish to express my deepest gratitude to my supervisor, Prof. Lihua Xie, for his professional guidance, kindly encouragement and unyielding support throughout my Ph.D. study. It was his incisive insights, thoughtful discussions, deep knowledge and outspoken suggestions that enabled me to forge ahead into my research. His attitudes towards research and life set a good example of what a scientist should be.

My special gratitude goes to Dr. Han Zou for not only leading me to explore research topics but also the sincere encouragement when I confronted kinds of difficulties. He is far beyond a collaborator but more like a mentor who selflessly shares life experiences to me.

Meanwhile, I would like to express my gratitude to my colleagues, Dr. Kexin Guo, Dr. Chen Wang, Prof. Hao Jiang, Dr. Zhenghua Chen, Dr. Yuxun Zhou, Dr. Keyu Chen, Yuecong Xu, Yuqing Gao, Shuxin Cao, Xu Fang, Pengfei Li, Dongzhe Wang, Prof. Zai Yang, Dr. Qianwen Xu, Dr. Kemi Ding, Dr. Zhirong Qiu, Dr. Yushen Long, Kun Cao, Muqing Cao, Dazhuo Wang, Haozhi Cao, Qi Li, Juncheng Li, Dr. Thien-Minh Nguyen, Hoang Thien and Dr. Abdul Hanif for their pleasant collaborations and fruitful discussions.

I need to thank my cool friends who locate around the world but make my Ph.D. study memorable and joyful. Special thanks go to Zhaoyang Zeng, Kai Wang, Jiangang Yang and Qin Wang for our unforgettable memories of competitions. Special thanks go to Yutian Shi, Wenkui Mi, Likuan Dong, Haobo Ma and Qiao Cui for their endless encouragement. Special thanks go to Shangyu Chen for sharing history and cool creeds. Special thanks go to Zhufeng Liu, Zhen Cao, Minhui Liang, Ruqi Zhou, Yingtian Wang and Fei Yu for our memories of gym, swimming, diving and werewolf. Special thanks go to Dr. Zuozhu Liu, Huaizheng Zhang, Nana Hou and Dr. Haiyan Yin for sharing academic and career stories.

Many thanks go to He Li, Dr. Yancheng Yuan and Yue Ai for sharing financial and start-up experiences.

Last but not least, I sincerely thank my beloved parents, for their unconditional love, warm encouragement and unwavering support. Without them never would I have been able to pursue the Ph.D. degree. Special thanks go to my grandfather, cousins and other family members. Love you all.

“Are we all lost stars, trying to light up the dark?”

—Lyrics of *Lost Stars*, Adam Levine

To my dear family

Abstract

Sensing and recognition technologies play an essential role in smart cities. Big data is acquired by various sensors deployed in almost every corner of our daily lives, which quantifies human activities and social systems. Smart sensing significantly facilitates people's lives and promotes the development of smart cities. Such success is due to not only the sufficient data by a variety of sensors but also the development of recognition technology driven by the progressive machine learning. As different sensors offer heterogeneous data, traditional statistical models may not be capable of tackling the high-dimensional data. To this end, neural networks with high fitting ability were proposed, which further evolved to deep neural networks as the neural layers become deeper and deeper for better capacity of feature extraction. Using deep learning, data of distinct sensors can be learned jointly for accurate recognition.

Nevertheless, towards robust smart sensing and recognition technology, there are still some bottlenecks that hinder the sensing performance and restrain the application scenarios. In this thesis, we propose systemic solutions and algorithms to overcome the longstanding challenges from two perspectives: sensing tools and recognition models.

From the perspective of sensing tools, different sensors are applied to their suitable scenarios. For instance, inertial Measurement Unit (IMU) is utilized for mobile devices and cameras take charge of surveillance. However, in smart homes, existing sensors have their intrinsic shortcomings such as the privacy issue of camera, the requirement of carrying devices for IMU and low granularity of data for Pyroelectric Infrared Sensor (PIR). To obtain a cost-effective, privacy-preserving and fine-grained sensing tools, this thesis pioneers a WiFi-based sensing platform which extracts Channel State Information (CSI) data directly from routers instead of specific network cards. CSI describes the propagation situation of wireless signals which can be affected by human motions. Therefore, these patterns of variant signals render WiFi-based sensing to be possible. The proposed platform is an

IoT-enabled platform which leverages off-the-shelf (COTS) WiFi routers for sensing. To demonstrate its effectiveness, we develop an occupancy detection system and a sedentary behavior monitoring system, and evaluate their performance in real-world scenarios. Traditional statistical features are employed for static and dynamic activity recognition. We further propose a novel approach for a more fine-grained recognition application — WiFi-based gesture recognition using the platform. Since gestures are too subtle to capture, many environmental noises are recorded in CSI samples. Extracting effective features for recognition from noisy data is a non-trivial task using deep neural networks. To this end, we propose a Siamese deep model with both spatial and temporal feature extractors, which discards the intrinsic noises of CSI data during feature learning. The proposed method also allows user to fine-tune the system using few samples, and thus is user-friendly.

From the perspective of recognition, this thesis focuses on dealing with the domain shift problem that hinders the generalization of recognition models. For instance, in WiFi-based activity recognition, the model trained in one environment cannot generalize well in a new environment due to the spatial shift existing in different environments. In visual sensing, images are captured by kinds of cameras with different brands and parameters, which hinders the performance of visual recognition models even more severely. To deal with these problems, domain adaptation minimizes the distribution discrepancy between source domain (*e.g.* the original environment for training) and target domain (*e.g.* the new environment for testing). This thesis contributes to four novel algorithms for real-world domain adaptation scenarios.

Firstly, we formulate the Edge Domain Adaptation (EDA) problem that addresses the domain shift between a model trained in center server and the model running in edge devices. Then MobileDA is proposed to deal with this problem by joint minimization of domain disparity and model complexity. It bridges the gap between domain adaptation algorithms and edge computing.

Secondly, current domain adaptation methods require careful manual hyper-parameter tuning, which is not realistic for unmanned systems such as smart cars. To stabilize the training of adversarial domain adaptation, Max-margin Domain-Adversarial Training (MDAT) is developed in this thesis to realize stable convergence without cumbersome manuals.

Thirdly, regarding the cross-domain feature learning, we further study the transferability and discriminability of deep features generated by domain adaptation models, which achieves a better performance due to the preservation of feature discriminability. This motivates us to propose Asymmetric Adversarial Domain Adaptation (AADA) that upgrades MDAT to a more generic framework. AADA also contributes to explaining the underlying mechanism behind MDAT.

Fourthly, Imbalanced Domain Adaptation (IDA) is proposed for the real-world imbalanced data since label shift commonly occurs to realistic data distribution. As the source domain and target domain may have imbalanced classes, IDA refers to the situation when covariate shift and label shift exist simultaneously. Then the proposed Cluster-level Discrepancy Minimization (CDM) conducts domain alignment after unsupervised clustering, which effectively aligns the domain shift and the label shift. Four algorithms for domain adaptation have been evaluated on public visual benchmark (*e.g.* Image-CLEF and Office-Home) and WiFi-based gesture recognition dataset. They achieve state-of-the-art performances as well as the specific goals such as training stability and better discriminant features.

In summary, the thesis contributes to robust sensing and recognition technology. For smart sensing, the WiFi-based sensing platform provides fine-grained CSI data, which can empower many applications including human activity recognition, gesture recognition, person identification, crowd counting and indoor localization. WiFi-based sensing is necessarily complementary to seamless sensing in smart homes or buildings. For robust recognition models, four domain adaptation algorithms are proposed in the thesis. They greatly give impetus to dealing with real-world domain adaptation scenarios such as long-tailed data or automatic model adaptation in edge devices. These algorithms can be further applied to improving the generalization abilities of deep models for WiFi-based sensing and visual sensing including object recognition, object detection, sentiment segmentation and person re-identification.

Contents

Acknowledgements	ix
Abstract	xiii
List of Figures	xxi
List of Tables	xxv
Symbols and Acronyms	xxvii
1 Introduction	1
1.1 Motivation and Objectives	1
1.2 Major Contributions	5
1.3 Organization of the Thesis	7
2 Literature Review	9
2.1 Overview of Sensing Techniques	9
2.1.1 Cameras	9
2.1.2 Pyroelectric Infrared (PIR)	10
2.1.3 Inertial Measurement Unit (IMU)	10
2.1.4 Radio Frequency (RF)	11
2.1.5 Lidar	11
2.1.6 Environment Sensors (ES)	11
2.1.7 Wireless Local Area Network (WiFi)	12
2.2 Overview of Learning Algorithms for WiFi-based Sensing	13
2.2.1 Traditional Statistical Learning	13
2.2.2 Deep Representation Learning	13
2.2.3 Few-shot Learning	15
2.2.4 Transfer Learning for Robust Feature Extraction	15
2.2.5 Cross-Modal Learning for Annotation Assistance	16
2.2.6 Multi-Modality Learning for Comprehensive Sensing	16
2.2.7 Model Compression for Edge Computing	16
2.3 Domain Adaptation Algorithms	17
2.3.1 Learning Theory of Domain Adaptation	18

2.3.2	Domain Adaptation by Global Statistics	18
2.3.3	Adversarial Domain Adaptation	20
2.3.4	Transferability and Discriminability	21
2.3.5	Covariate Shift and Label Shift	22
2.3.6	Other Domain Adaptation Categories	23
2.3.7	Domain Adaptation Benchmarks	24
2.4	Conclusion	26
I	Smart Sensing	27
3	WiFi-based IoT-enabled Sensing Platform	29
3.1	Introduction	29
3.2	Preliminary: Channel State Information	32
3.3	Agile WiFi-enabled IoT Platform for WiFi-based Sensing	33
3.3.1	IoT Platform	34
3.3.2	Cloud Server and User Terminal	36
3.4	Case Study: Coarse Sensing for Occupancy Detection	37
3.4.1	Insights: How Occupancy Affects Wireless Signals	37
3.4.2	Methodology	39
3.5	Case Study: Fine-grained Sensing for Sedentary Behavior Monitoring	40
3.5.1	Sedentary Behavior	40
3.5.2	How Sedentary Activities Affect CSI	42
3.5.2.1	Activities in sedentary behavior	43
3.5.2.2	Connections of Activities and WiFi data	43
3.5.3	Methodology	46
3.5.3.1	System Overview	46
3.5.3.2	Coarse Recognition by Motion Extraction	47
3.5.3.3	Data Preprocessing	49
3.5.3.4	Dynamic Activity Recognition	50
3.5.3.5	Static Posture Identification	50
3.5.3.6	Profile Update and Close Loop	52
3.6	Evaluation of Occupancy Detection	52
3.7	Evaluation of Human Activity Recognition	54
3.7.1	Overall Evaluation	55
3.7.2	Impact of LOS/NLOS Circumstances	56
3.7.3	Impact of Environmental Dynamics and People Around	57
3.7.4	Comparison of Rough and Fine-grained CSI	58
3.8	Conclusion	59
4	Robust Representation Learning for Few-shot WiFi-based Gesture Recognition	61
4.1	Introduction	61
4.2	System Design	64

4.2.1	IoT Platform for Sensing Gestures	64
4.2.2	Insights: Learning Gestures from CSI Phase	64
4.3	A Siamese Recurrent Convolutional Network Architecture	65
4.3.1	Model	65
4.3.2	Transferable Pairwise Loss	67
4.4	Experiment	70
4.4.1	System Settings and Data Collection	70
4.4.2	Learning Configurations	71
4.4.3	Evaluation	73
4.5	Conclusion	77
II	Domain Adaptation for Robust Recognition	79
5	Efficient Domain Adaptation for Transfer Learning in the Edge	81
5.1	Introduction	81
5.2	Method	85
5.2.1	Problem Definition and Notation	85
5.2.2	Pre-training Teacher on Source Domain	85
5.2.3	Distillation: Target Knowledge Transfer	86
5.2.4	Adaptation: Diminishing Domain Disparity	87
5.3	Experimental Results	89
5.3.1	Datasets	89
5.3.2	Setup and Implementation Details	90
5.3.3	Results	91
5.3.4	Ablation Study	94
5.3.5	Analysis	96
5.4	Conclusion	98
6	Stable and Comprehensive Domain Adaptation: Max-margin Domain-Adversarial Training	99
6.1	Introduction	99
6.2	Problem Formulation	101
6.2.1	Problem Definition and Notations	101
6.2.2	Unbalanced Minimax Game in Domain-Adversarial Training	102
6.2.3	Max-margin Domain-Adversarial Training	103
6.2.4	Optimal Solution of MDAT	104
6.2.5	Connection to Domain Adaptation Theories	106
6.2.6	Merits: MDAT vs. DAT	107
6.3	Experiment	108
6.3.1	Setup	108
6.3.2	Overall Results	109
6.3.3	Analyses	111
6.4	Conclusion	114

7	Generalized MDAT: Asymmetric Adversarial Domain Adaptation	115
7.1	Introduction	115
7.2	Revisiting Transferability and Discriminability in Symmetric Adversarial Domain Adaptation	118
7.2.1	The Theory of Domain Adaptation	118
7.2.2	Limitations and Insights	119
7.3	Asymmetric Adversarial Domain Adaptation	121
7.3.1	The Learning Framework	121
7.3.2	Discussions and Theories	123
7.4	Experiments	124
7.4.1	Experimental Setup	124
7.4.2	Overall Results	125
7.4.3	Spectral Analysis using SVD	128
7.4.4	Analytics and Discussions	129
7.5	Conclusion	132
8	Imbalanced Domain Adaptation for Long-tailed Data	135
8.1	Introduction	135
8.2	Imbalanced Domain Adaptation	137
8.2.1	Methodology Overview	138
8.2.2	Discriminative Cluster Learning	139
8.2.3	Cluster-level Discrepancy Minimization	140
8.2.4	Boundary Refinement	142
8.2.5	Theoretical Insight	143
8.3	Experiment	144
8.3.1	Setup	144
8.3.2	Summary of Results	146
8.3.3	Analysis	148
8.4	Conclusion	151
9	Conclusion and Future Work	153
9.1	Conclusions	153
9.2	Future Works	155
9.2.1	Machine Learning	156
9.2.2	Artificial Intelligence of Things (AIoT)	157
	List of Author's Awards, Patents, and Publications	159
	Bibliography	167

List of Figures

1.1	The structure of the thesis.	4
2.1	The architecture of LeNet-5. (Figure source: Lecun et al. [1])	14
2.2	The example images of Office-Home.	25
3.1	Overview of the system architecture.	34
3.2	Curves of CSI subcarrier values from unoccupied (U) to occupied (O) scenarios. We can see that there exists obvious difference of curves between O and U.	38
3.3	STI values of one TX-RX pair. The occupants appear for four times indicated by four peaks.	38
3.4	Multi-path effect in wireless signal propagation and various daily activities in sedentary behavior.	41
3.5	Different trends of subcarrier clusters in CSI stream.	42
3.6	The CSI sequences of various dynamic activities.	45
3.7	Subcarrier difference among static activities.	45
3.8	Framework of CareFi.	46
3.9	Noise of each subcarrier follows a Gaussian distribution independently.	48
3.10	Comparison of raw and denoising data on subcarrier #20 in CSI stream.	48
3.11	Top 4 principal components of CSI using PCA.	51
3.12	The layouts of the dormitory and office room.	52
3.13	ROC curves of CareFi in LOS/NLOS situations.	55
3.14	Confusion matrix of dynamic activity recognition.	56
3.15	Confusion matrix of posture identification in LOS and NLOS situations.	57
3.16	Comparison of posture identification accuracy in LOS and NLOS situations.	58
3.17	Overall performance of posture identification with or without profile update.	58
4.1	Illustration of our Siamese recurrent convolutional architecture. The model accepts a new gesture input and extracts its deep features, which are then compared with all the historical one-shot templates to choose the most similar one.	63

4.2	Phase difference of CSI readings, including two kinds of gestures, in which the first two are <i>downward</i> and the last two are <i>leftward</i> . . .	65
4.3	How pairwise loss and domain confusion loss work for biased data. Three colors represent different domains and the shapes denote their labels.	68
4.4	Six gestures used in the experiment.	71
4.5	The environmental settings of the meeting rooms.	71
4.6	Confusion matrix of gesture recognition using different methods. . .	72
4.7	Test sample clustering results.	75
4.8	Confusion matrix of granularity test. We can see that hand-level gestures are easily classified while finger-level ones are tough to deal with due to the granularity of CSI data.	76
5.1	We transfer the novel knowledge from a sophisticated teacher model trained in the server to a simple student model running in the edge side while considering the domain shift in the new environment such as smart homes.	82
5.2	Illustration of the proposed approach.	83
5.3	Convergence and domain discrepancy. (a)(b) The loss and test error during training for Deep-CORAL and our MobileDA on $\mathbf{I} \rightarrow \mathbf{C}$ (<i>ImageCLEF-DA</i>). (c) $\mathcal{A}$ -distance on <i>Office-31</i> with transfer tasks of $\mathbf{A} \rightarrow \mathbf{W}$ and $\mathbf{W} \rightarrow \mathbf{D}$	94
5.4	Confusion matrix of source-only AlexNet (70.8%), Deep-CORAL (73.3%) and MobileDA (89.5%) on the transfer task of $\mathbf{P} \rightarrow \mathbf{I}$ on <i>ImageCLEF-DA</i>	95
5.5	t-SNE plots of MobileDA features on the transfer task of $\mathbf{P} \rightarrow \mathbf{I}$ on <i>ImageCLEF-DA</i>	95
6.1	The proposed architecture is composed of a shared feature extractor G_e for two domains, a label predictor G_y and a reconstruction network G_r . In addition to the basic supervised learning in the source domain, our adversarial reconstruction training enables the extractor G_e to learn domain-invariant features. Specifically, the network G_r aims to reconstruct the source samples x^s and to impede the reconstruction of the target samples x^t , while the extractor G_e tries to fool the reconstruction network in order to reconstruct the target samples x^t	101
6.2	The <i>inter-twining moons</i> toy problem. Red and green dots indicate source samples, while blue dots are target samples. Black lines indicate the changes of decision boundaries during 10 epochs training.	109
6.3	The training procedure <i>w.r.t.</i> loss and test accuracy. $\mathcal{L}_e$ is the training loss of reconstructing target samples in Eq. 6.6. $\mathcal{L}_r$ is the training loss of reconstruction network in Eq. 6.4. $\mathcal{L}_d$ is the domain loss in DAT [2]. α is the penalty term of $\mathcal{L}_e$ and $\mathcal{L}_d$ in MDAT and DAT, respectively.	111

6.4	T-SNE visualization on SVHN → MNIST with their corresponding domain labels (red: target; blue: source) and category labels (10 classes) shown in the left and right subfigures, respectively.	113
7.1	Comparison between previous Domain Adversarial Training (DAT) method and ours. Left: The discrepancy region is large before adaptation. Middle: DAT based on a domain classifier aligns two domains together and pushes them as close as possible, which hurts the feature discriminability in the target domain. Right: AADA fixes the source domain, regarded as the low-energy space, and pushes the target domain to approach the source one, which makes use of the well-trained classifier in the labeled source domain.	117
7.2	AADA model constitutes a shared feature extractor G_f , a classifier G_y and an autoencoder G_a . Except for supervised learning on the source domain, the autoencoder plays a domain discriminator role that learns to embed the source features and push the target features away, while the feature extractor learns to generate target features that can deceive the autoencoder. Such process is an asymmetric adversarial game that pushes the target domain to the source domain in the feature space.	121
7.3	Measures of discriminability and transferability.	129
7.4	The training stability and the sensitivity of the hyper-parameters.	130
7.5	Accuracy by varying m (left) and γ (right).	130
7.6	The t-SNE visualization of the embeddings z on task SVHN → MNIST . The top figures are visualization with domain labels (blue: source, red: target). The bottom figures are visualization with category labels (10 classes).	132
8.1	(a) Standard domain adaptation assumes that the source domain and the target domain share the same label distributions, which is followed by prevailing UDA benchmarks with balanced category. (b) Imbalanced domain adaptation considers the real-world situation when label shift exists. For example, if the source domain is an urban area there are more cars and buses, while if the target domain is a rural area there are more bikes and horses.	136
8.2	The illustration of existing UDA solutions and our proposed CDM for IDA problem. Existing approaches cannot deal with label shift by direct minimization. Our approach learns discriminative clusters and then minimizes the cluster-level discrepancy with boundary refinement, overcoming the negative effect of label shift <i>w.r.t.</i> the discrepancy minimization.	137
8.3	Dataset protocols for evaluation and real-world dataset SVHN.	145
8.4	The confusion matrix on A→W using U2U protocol.	147
8.5	(a) The accuracy on A→W with different level of label shift. (b)(c) The comparison of average accuracy on <i>Image-CLEF</i> based on ResNet18.	149

- 8.6 The t-SNE visualization of DAN [3] and the proposed CDM on $P \rightarrow I$ (*Image-CLEF*). The left figure is with the domain label (red: source; blue: target) while the right figure is with the category label (in different colors). 150

List of Tables

2.1	Overview of major sensing technologies	9
3.1	Occupancy Detection Performance	54
4.1	Overall accuracy of different methods	72
4.2	Average Accuracy of Cross-domain Evaluation	76
5.1	Comparisons on WiFi Gesture Recognition.	91
5.2	Accuracy (mean) of unsupervised domain adaptation on <i>Office-31</i> datasets across 3 independent runs.	92
5.3	Accuracy (mean) of unsupervised domain adaptation on <i>ImageCLEF-DA</i> datasets across 3 independent runs.	92
5.4	Accuracy (mean) of unsupervised domain adaptation on <i>Office-Home</i> datasets across 3 independent runs.	93
5.5	Ablation study on <i>ImageCLEF-DA</i>	93
5.6	Inference time on IoT devices	94
5.7	Accuracy on <i>ImageCLEF-DA</i> (I→P) <i>w.r.t</i> λ	96
6.1	The network architecture used in the <i>Digits</i> experiments.	109
6.2	We compare with general, statistics-based (S), reconstruction-based (R) and adversarial-based (A) domain adaptation approaches. We repeated each experiment for 3 times and report the average and standard deviation (std) of the test accuracy in the target domain.	110
6.3	Comparisons on WiFi gesture recognition.	110
6.4	Accuracy (mean) of unsupervised domain adaptation on <i>Office-Home</i> datasets across 3 independent runs.	110
6.5	Accuracy (%) on SVHN → MNIST	112
6.6	Visualizing the source image, target images and reconstructed target images (R-Target Images) for four digit adaptation tasks.	112
7.1	The network architecture used in digit adaptation	125
7.2	The network architecture used in Office-Home and VisDA-17.	126
7.3	Accuracy (mean±std) on five <i>Digit</i> benchmarks for domain adaptation across 3 independent runs.	126
7.4	Accuracy (mean) of unsupervised domain adaptation on <i>Image-CLEF</i> datasets across 3 independent runs.	127

7.5	Accuracy (mean) of unsupervised domain adaptation on <i>Office-Home</i> datasets across 3 independent runs.	127
7.6	Accuracy (%) on VisDA-2017 for domain adaptation.	128
8.1	The average classification accuracy (%) with the standard deviation on <i>Office-31</i> . For each column, left and right results are obtained by applying B2U and U2U protocol, respectively.	147
8.2	The average classification accuracy (%) with the standard deviation on <i>Image-CLEF</i> . For each column, left and right results are obtained by applying B2U and U2U protocol, respectively.	148
8.3	The average classification accuracy (%) with the standard deviation on <i>Office-Home</i> . For each column, left and right results are obtained by applying B2U and U2U protocol, respectively.	148
8.4	The classification accuracy (%) for ablation study on <i>Office-31</i> and <i>Office-Home</i> following B2U and U2U protocols.	149

Symbols and Acronyms

Symbols

$\mathbb{C}$	set of complex numbers
$\mathbb{R}$	set of real numbers
$\mathbb{R}^N$	set of real N dimensional vectors
$\mathbb{R}^{M \times N}$	set of real $M \times N$ dimensional matrices
$\mathcal{H}$	the hypothesis space
$\mathcal{H} \triangle \mathcal{H}$	the symmetric difference hypothesis space for a hypothesis space $\mathcal{H}$
$\mathcal{H}_k$	the reproducing kernel Hilbert space
$\ \cdot\ _2$	the 2-norm of a vector or matrix in Euclidean space
$\ \cdot\ _F$	the Frobenius norm of a matrix
$\ \cdot\ $	the amplitude of a complex number
$\angle H$	the angle of a complex number
$\mathbb{E}[X]$	mathematical expectation of a random variable X
$\hat{x}$	the empirical estimation of x
$\bar{x}$	the mean of a vector x
$\mathbf{1}$	all-ones column vector with proper dimension
$\int_{\Omega}$	the integrals over a domain Ω
$\triangleq$	the definition of a symbol
$Pr[X]$	a probability measure of a random variable X
$p(x)$	a marginal probability distribution of x
$\odot$	the Hadamard (component-wise) product
$\mathbb{I}[\cdot]$	the indicator function
$\sup \cdot $	the supremum of a formula
$[\cdot]^+$	$\max(0, \cdot)$
$\langle \cdot, \cdot \rangle$	the inner product of two vectors
$\text{cov}(\cdot, \cdot)$	the cross-covariance of two random variables

$G(\cdot; \theta)$	a function G parameterized by θ
$f \circ g$	the functions compositions $f(g(x))$
$d_k(\cdot, \cdot)$	the RKHS distance of two embeddings of distributions
$d_{JS}(\cdot, \cdot)$	the JensenShannon divergence of two distributions
$\epsilon_{\mathbb{Q}}(h)$	the expected error of a hypothesis h according to the distribution $\mathbb{Q}$
$x \sim \mathbb{Q}$	a sample x follows the distribution $\mathbb{Q}$

Acronyms

IoT	Internet of Things
CCTV	Closed-Circuit TeleVision
IMU	Inertial Measurement Unit
RF	Radio Frequency
SVM	Support Vector Machine
DNN	Deep Neural Networks
UDA	Unsupervised Domain Adaptation
ADA	Adversarial Domain Adaptation
SDA	Supervised Domain Adaptation
CSI	Channel State Information
CNN	Convolutional Neural Network
RNN	Recurrent Neural Network
LSTM	Long Short-Term Memory
DAT	Domain-Adversarial Training
MDAT	Max-margin Domain-Adversarial Training
AADA	Asymmetric Adversarial Domain Adaptation
EDA	Edge Domain Adaptation
IDA	Imbalanced Domain Adaptation
PIR	Pyroelectric Infrared Sensor
RFID	Radio Frequency Identification
KNN	K-Nearest Neighbours
ES	Environment Sensors
LDA	Linear Discriminant Analysis
WiFi	Wireless Local Area Network
RSSI	Received Signal Strength Indicator

PCA	Principal Component Analysis
DTW	Dynamic Time Warping
SVD	Singular Value Decomposition
MMD	Maximum Mean Discrepancy
MK-MMD	Multiple Kernel variant of MMD
RKHS	Reproducing Kernel Hilbert Space
GAN	Generative Adversarial Networks
GRL	Gradient-Reverse Layer
OFDM	Orthogonal Frequency Division Multiplexing
MIMO	Multiple Input Multiple Output
MQTT	MQ Telemetry Transport
QoS	Quality of Service
BMS	Building Management System
STI	Signal Tendency Index
SB	Sedentary Behavior
CV	Computer Vision
LOS	Line-Of-Sight
NLOS	Non-Line-Of-Sight
COTS	Commercial Off-The-Shelf
KDE	Kernel Density Estimation
FFT	Fast Fourier Transform
IFFT	Inverse Fast Fourier Transform
CIR	Channel Impulse Response
PPMCC	Pearson Product-Moment Correlation Coefficient
ROC	Receiver Operating Characteristic
HCI	Human-Computer Interaction
SGD	Stochastic Gradient Descent
FLOPs	Floating Point Operations
FC	Fully Connected layers
ARN	Adversarial Reconstruction Network
CDM	Cluster-level Discrepancy Minimization
MAML	Model-Agnostic Meta-Learning
NAS	Neural Architecture Search
GNN	Graph Neural Network
AIoT	Artificial Intelligence of Things

KG	Knowledge Graph
----	-----------------

Chapter 1

Introduction

1.1 Motivation and Objectives

The emergence of smart homes and smart cities is attributed to the rapid expansion of sensing and recognition technologies. Smart sensors empower the informatization of people's daily lives and artificial systems such as closed-circuit television (CCTV) for public area and mass traffic. Relying on the rising technology of artificial intelligence, we can further analyze the big data obtained by smart sensors to generate human-defined descriptions. One example is visual object recognition that outputs the desired results based on human knowledge and cameras. These results help smart cities deal with congestion and energy waste, or assist human to improve energy efficiency and security in smart homes. According to a recent market research, the Internet of Things (IoT) sensor market is expected to rise from USD 11.91 billion in 2019 to USD 42.67 billion by 2025 [4]. As IoT sensors play the most crucial role in every aspect of automation, they will significantly promote the Industry 4.0 initiatives, healthcare for elderly monitoring and autonomous driving.

In the past few decades, various IoT sensors, such as Inertial Measurement Unit (IMU), Radio Frequency (RF) sensor, infrared sensor, environmental sensors, radar and camera, have been exploited for different smart sensing applications. IMU is widely used in mobile devices for human activity recognition and step counting, but the users need to equip with IMUs, which is not convenient for some situations [5]. RF sensor can detect human activities in a device-free manner, but this requires a large scale of deployment of many RF sensors [6]. Infrared sensor is

commonly seen in occupancy detection devices in indoor environments [7], but it has very low granularity and cannot detect relatively static human motions. Environment sensors [8], including CO2 sensor, temperature sensor, humidity sensor, air quality sensor and wind sensor, play an essential role in smart buildings. They help improve the indoor comfort level. Lidar provides very accurate data of point cloud, which is quite expensive but makes it suitable for autonomous driving and robotics. Camera, normally utilized for CCTV, can be seen almost everywhere in the public areas. However, in smart homes, it arouses privacy issue and can be affected easily by low illumination or occlusion. As such, when we focus on smart homes, a device-free, cost-effective and fine-grained sensing technique is demanded. In this thesis, we explore a WiFi-based sensing platform that leverages existing WiFi access points for accurate sensing in Chapter 3.

Sensor constitutes the first step of sensing technique as sensing data is only the raw materials. How to process these materials and produce the desired outputs is even more important. For simple applications such as occupancy detection, a threshold might be sufficient. However, for hard tasks that deal with high-dimensional data such as visual data, statistical learning offers a powerful tool to model sensor data in a data-driven fashion. Many hard-to-model patterns can be captured by statistical learning and well classified by optimal classifiers such as Support Vector Machine (SVM) [9]. To achieve better fitting ability of statistical model, artificial neural networks were proposed, and they went deeper and deeper recently for high-level tasks that even approach human levels of sensing and recognition. Deep Neural Networks (DNNs) have been exploited especially in computer vision tasks such as object recognition, semantic segmentation and image captioning [10]. DNNs automatically learn robust representations of sensing data by back-propagation [11]. Since humans cannot understand and annotate sensing data in pixel level, the representation learning of DNNs assists us to find robust patterns for better understanding and recognition models for various applications. In this thesis, we explore new algorithms for WiFi-based sensing for real-world scenarios in Chapter 3 and 4.

Using deep learning in smart sensing, we encounter practical problems due to environmental dynamics. For instance, we train a WiFi-based human activity recognition model in one room. When applying this model to a new environment, its performance decreases a lot. The reason is that data-driven model may learn

environment-specific information that is irrelevant to human activities and hinders the generalization ability. In visual recognition, such situation becomes more often. An example is that most of the training images in public dataset are collected in sunny or cloudy weather but the testing scenario might be snowy, rainy or even hazy weather. The environmental factors change the data distribution of images, which is defined as domain shift or covariant shift in statistical learning [12]. Such shift consists of pixel-level shift such as blur or hazy noise and feature-level shift such as object style or imaging style. To bridge the domain shift, domain adaptation refers to transferring knowledge from a label-rich source domain to a label-scarce or even unlabeled target domain, which adapts the model to the new environment using few annotations in target domain. Current domain adaptation approaches mainly rely on minimization of statistical measures or adversarial training. As Adversarial Domain Adaptation (ADA) yields excellent performance, it has drawn much attention. In this thesis, to enable a more robust and generic recognition model in smart sensing, we focus on computation complexity issue, stable adversarial training issue, feature discriminability issue and imbalanced long-tailed data for realistic domain adaptation algorithms.

The structure of this thesis is illustrated in Figure 1.1, in which the core concept is robust smart sensing. To this end, in Part I, we pioneer a new sensing platform – an WiFi-enabled IoT sensing platform and propose multiple novel recognition models for different scenarios and applications. Then we find that the domain shift severely degrades the performance of statistical learning model for WiFi-based and visual sensing. Therefore, in Part II, we explore to bridge the gap between the ideal training circumstance and the real-world scenarios with environmental dynamics by developing domain adaptation (i.e. transfer learning) algorithms in order to achieve robust recognition in smart sensing. The algorithms proposed in Part II contribute to robust smart sensing technology such as WiFi-based gesture recognition in Part I.

In summary, the objective of this thesis is to empower robust sensing and recognition by means of statistical learning and transfer learning (*e.g.* deep learning and domain adaptation, respectively).

FIGURE 1.1: The structure of the thesis.

1.2 Major Contributions

Our main contributions can be stated as follows:

WiFi-based Sensing Platform and Its Applications. Existing WiFi-based sensing platform leverages Intel 5300 NIC and only runs on mini-PC. In this thesis, we design a WiFi-enabled IoT system that extracts fine-grained Channel State Information (CSI) and runs on both mini-PC and WiFi routers. Furthermore, for different granularity of sensing, the platform is designed to conduct the computations in either cloud or edge side. To the best of our knowledge, it is the first platform which completely integrates CSI-based WiFi devices, cloud server and user end. It makes a good trade-off between performance and availability. To corroborate its capacity, we develop an occupancy detection method based on similarity measurement of CSI curves and a sedentary behavior monitoring system based on statistical feature learning. We evaluate these applications by real-world experiments to demonstrate their accuracy, robustness and scalability.

WiFi-based Gesture Recognition via Siamese network. Though the proposed platform extracts delicate CSI data, utilizing it for gesture recognition is still difficult since the CSI data is never clean in real-world settings. The measured wireless signal could be heavily distorted by walls, object placement and individual heterogeneity, which reduces the effectiveness of data-driven deep models due to overfitting. Moreover, gesture recognition in a new environment suffers from domain shift, and we only allow very few annotations by users. To this end, we propose a Siamese architecture with transferable pairwise loss to clean up other structured but irrelevant features such as environmental dynamics. In this manner, only the features that help identify the gestures are learned by our network. Convolutional Neural Network (CNN) and Long Short-Term Memory (LSTM) are leveraged for spatial and temporal pattern extraction, respectively. We conduct real-world experiments in multiple environments and the results show significant improvement against existing systems.

Efficient Domain Adaptation in the Edge. Existing domain adaptation tackles the problem of domain shift that may exist between the training data in a center server and the testing data in the edge. However, as the edge resources cannot satisfy the real-time computation of deep model, the computational bias between these two sides should also be considered. To this end, we propose MobileDA

that simultaneously learns a robust domain-invariant representation and simplifies the network architecture for efficient inference of deep models. MobileDA reduces the domain shift by statistical metrics and further enhances the deep model by an innovative cross-domain distillation. Extensive experiments are conducted on IoT-based and visual benchmarks. MobileDA achieves both better results and low latency of computations.

Max-margin Domain-Adversarial Training. In domain adaptation, adversarial domain adaptation is very attractive since it shows remarkable performance on most benchmarks. Nevertheless, ADA cannot be trained in a stable manner, which requires manual hyper-parameter tuning and thus hinders its availability for real-world domain adaptation solution. To stabilize ADA, we propose a new ADA approach, namely Max-Margin Domain-Adversarial Training (MDAT), that balances the adversarial game by a max-margin loss. MDAT achieves better pixel-level alignment and has a stable training procedure. In principle, MDAT is generic and can be used to enhance any domain adaptation methods that leverage domain alignment as an ingredient. In visual and WiFi domain adaptation benchmarks, it achieves tremendous improvement compared to conventional Domain-Adversarial Training (DAT).

Asymmetric Adversarial Domain Adaptation. In MDAT, to stabilize DAT, we replace the domain classifier with a reconstruction network. We further exploit the functionality of such domain classifier and generalize the idea of MDAT to Asymmetric Adversarial Domain Adaptation (AADA). With similar formulations, we can replace the domain classifier with any energy-based functions which map the source domain to the low-energy space and enforces the target domain to approach it. In this fashion, feature learning in the source domain is not deteriorated by adversarial training. We apply spectral analysis to demonstrate better discriminability of features learned by AADA compared to DAT. Comprehensive experiments show that AADA outperforms DAT significantly on all visual benchmarks.

Cluster-level Discrepancy Minimization for Imbalanced Domain Adaptation. In realistic problems, apart from domain shifts, long-tailed data and label shifts are very common. Existing domain adaptation based on feature alignment cannot tackle these problems, which motivates us to propose Cluster-level Discrepancy Minimization. Similarity learning is firstly conducted to obtain tight clusters,

and then we align the prototypes of these clusters from different domains, which eliminates the negative effect of imbalanced label distribution and enhances feature learning for long-tailed data. To evaluate it, we propose two new protocols for testing by splitting public datasets following Pareto distribution. We conduct extensive experiments on three visual datasets following our protocols, and demonstrate that CDM performs state-of-the-art results for all settings.

1.3 Organization of the Thesis

The remainder of this thesis is organized as follows:

Chapter 2 presents a literature review on various sensing technologies and recognition algorithms. Particularly, WiFi-based sensing and its applications are reviewed, which is closely related to this thesis. Then the learning theory and recent approaches of domain adaptation are also reviewed in this chapter. Some challenges of these methods when applying them to realistic problems are pointed out, which inspires this research.

The thesis is divided into two parts: smart sensing and domain adaptation for robust recognition. Part I, which consists of Chapter 3 and 4, mainly introduces the contribution on WiFi-based sensing. Part II, composed of Chapter 5, 6, 7 and 8, illustrates four domain adaptation algorithms for deep learning models and real-world application scenarios.

Part I - Smart Sensing: this part exploits the WiFi-based sensing technology from the perspectives of sensing platform and methodology. Several applications are empowered and evaluated in the real-world scenarios.

Chapter 3 proposes a novel WiFi-based sensing platform for robust sensing and efficient computing, which extracts fine-grained CSI data and can be deployed by embedded devices in a large-scale manner. In this chapter, two case studies of this system are proposed including occupancy detection and sedentary behavior monitoring. The investigation of how CSI data is related to human actions is detailed, and then follows by the algorithm designs of two applications.

Chapter 4 proposes a Siamese architecture for WiFi-based gesture recognition with only limited data in the new environment, which aims to capture the most useful information hidden in the CSI data with environmental noises.

Part II - Domain Adaptation for Robust Recognition: to address the performance degradation caused by spatial and temporal dynamics for statistical learning model, we propose four transfer learning algorithms that enables the statistical model to adapt to the new environment in an unsupervised manner. These algorithms render the recognition model to learn domain-invariant features under several realistic circumstances, which is effective for both WiFi-based and vision-based sensing.

Chapter 5 proposes a new problem, namely Edge Domain Adaptation (EDA), which refers to domain adaptation in the edge devices. EDA aims to train a computational-efficient model that deals with domain shift, and thus it can be quite effective in smart factories.

Chapter 6 proposes a new domain alignment approach, namely max-margin domain-adversarial training, which aims to stabilize the training procedure for domain adaptation in an automatic manner. Theoretical justifications are provided to demonstrate how MDAT reaches a Nash Equilibrium point when feature distributions are well aligned.

Chapter 7 proposes asymmetric adversarial domain adaptation that generalizes the method in Chapter 6. From the perspective of feature discriminability and transferability, Chapter 7 explains how MDAT and AADA improve the performance of domain adaptation. Some quantitative results are discussed with respect to several public benchmarks.

Chapter 8 proposes a realistic domain adaptation scenario, namely Imbalanced Domain Adaptation (IDA), which refers to domain adaptation under both covariate shift and label shift. In this chapter, cluster-level discrepancy minimization is presented to deal with the IDA problem. Two protocols for evaluations are proposed and extensive experiments are also detailed and discussed.

Chapter 9 concludes the thesis and envisions some potential future research.

Chapter 2

Literature Review

2.1 Overview of Sensing Techniques

Various sensing technologies have been developed to provide different granularity of sensing solutions for many applications. The main characteristics of these technologies (*e.g.* the requirement of extra devices on the user side, cost, granularity and the preservation of privacy) are summarized in Table 2.1. These technologies and their recent progresses are detailed in this section.

2.1.1 Cameras

Nowadays, cameras are everywhere as CCTV is widely deployed for security. Massive images and videos can be captured via cameras, and with the development of

TABLE 2.1: Overview of major sensing technologies

Technology	Granularity	Device-free	Cost	Privacy-preserving
Camera	High	Yes	Moderate	No
Infrared	Low	Yes	Low	Yes
IMU	Middle	No	Low	Yes
RF	Middle	Yes	Moderate	Yes
Radar	High	Yes	High	No
ES	Low	Yes	Low	Yes
WiFi	High	Yes	Low	Yes

deep learning, many high-level tasks are enabled such as object recognition, object detection and semantic segmentation [10]. Visual sensing has high granularity of sensing materials, but also arouses privacy concerns such as sensing in hospitals or smart homes. Furthermore, illumination and occlusion also severely affect the performance of visual sensing under some specific circumstances [13], which requires other sensing techniques for complement.

2.1.2 Pyroelectric Infrared (PIR)

Pyroelectric infrared sensor (PIR) can detect human motions in a designated area. A surface electric charge is generated by a PIR sensor when it is exposed to heat in the form of infrared radiation. The range of infrared wavelength is from 700 nm to 1mm and this is longer than visible light, which makes PIR-based sensing less intrusive than visual sensing. PIR is widely used in building occupancy detection [7] and intruder detection [14]. However, the sensing granularity of PIR is not high, so it cannot be employed for fine-grained activity recognition or other high-level tasks.

2.1.3 Inertial Measurement Unit (IMU)

IMU has the capacity of measuring the velocity, gravitational forces and orientation of moving by gyroscopes and accelerometers. It is widely equipped in mobile phones and smart watches that apply IMU to step counting. The IMU data can also be utilized for human activity recognition. Since human activities are composed of a series of motions that are easily captured by IMU, statistical learning model is able to extract activity patterns from these temporal data [15]. IMU-based activity recognition methods [16] can identify simple activities including walking, jogging, sitting, standing, etc. It is also noteworthy that attaching IMU to different parts of body leads to distinct measurements for the same activity. This issue can be partially overcome by transfer learning [17]. The drawback of IMU is that it requires users to equip it and thus IMU cannot work in many situations when users do not carry any devices.

2.1.4 Radio Frequency (RF)

Radio Frequency Identification (RFID) belongs to radio based sensing [18]. Other radio based methods include ZigBee radio and WiFi radio. Human activities are captured by RF-signals in a passive manner. Traditional statistical learning models such as K-Nearest Neighbours (KNN) achieve satisfactory performance for RF-based activity recognition. Deploying multiple RF sensors enables device-free recognition of simultaneously conducted activities [6]. The disadvantage of RF-sensors is that the sensing range is rather limited so RF sensors should be used in a large scale, which may lead to higher cost and additional deployment process.

2.1.5 Lidar

Lidar is capable of measuring distances to a target by illuminating the target with laser light and then measuring the reflected light. Using laser return times and wavelengths, 3D representations of the target can be obtained. Its sensing granularity is very high, which enables lidar-based object recognition [19] and object segmentation [20]. As deep learning techniques provide powerful modeling ability of Lidar 3D point cloud data, Lidar plays an essential role in autonomous driving for accurate sensing of surrounding environment [21]. Lidar is expensive and the processing of its data requires high computing resources, which makes it not suitable for lightweight sensing applications such as human activity recognition in smart homes.

2.1.6 Environment Sensors (ES)

Environment sensors mainly include light, wind, air quality, temperature, humidity and CO₂ sensors. Being deployed them in indoor environment, they usually reflect human living comfort. ES are usually utilized for improving thermal comfort and saving energy [22]. The data collected by ES is also used to predict the occupancy number [23]. Naturally, as more people lead to more heat and CO₂ concentration, simple statistical learning methods such as Linear Discriminant Analysis (LDA) and Random Forest (RF) can produce satisfactory results. The drawbacks of ES is that the large-scale deployment of ES brings high cost and ES cannot provide high-resolutional data for fine-grained sensing.

2.1.7 Wireless Local Area Network (WiFi)

Nowadays, WiFi access points have been deployed in most of the commercial and residual buildings, and nearby every IoT devices, such as sound bar, TV, thermostats and power switch, are equipped with WiFi module. To attain a better trade-off between accuracy and cost, WiFi-based methods come into existence. Received Signal Strength Indicator (RSSI) is firstly exploited but it is only a rough scalar that indicates the quality of signal strengths. Recently, Channel State Information (CSI) is extracted from the physical layer of wireless communications, which reflects the situations of the multi-path propagation of wireless signals in indoor environments. WiFi-based sensing is conducted in a device-free manner, and the granularity of CSI data is moderately high.

For the CSI extraction tool, Intel 5300 NIC [24] is the first tool to extract CSI from the network card. However, it only extracts a sampled version of complete CSI data, *i.e.* 30 subcarriers CSI of the complete 56 subcarriers in 20 MHz bandwidth. Then Atheros Tool [25] is developed to extract the complete CSI data in 20/40 MHz bandwidth. Theoretically, it enables all Atheros network cards to capture CSI data. However, both Intel 5300 NIC and Atheros Tool are only communication-level tools that focus on extracting CSI data. For IoT-based applications, a generic WiFi sensing platform is needed with the CSI tool equipped, which motivates us to design our IoT-enabled WiFi sensing platform.

Recent literature has witnessed many successful employments of CSI measurements for various applications, such as device-free indoor localization [26], human activity recognition [27, 28], human identification [29], smoking detection [30] and crowd counting [31, 32]. E-eyes [33] employs the similarity of CSI waves and dynamic time warping to recognize human activities. CARM [27] models the human walking speed and CSI, which improves the recognition model. WiKey [34] recognizes the keystroke by means of principal component analysis (PCA) and dynamic time warping (DTW). With the development of deep learning, deep models are introduced for CSI-based sensing, achieving excellent results [35–38]. However, due to the individual heterogeneity and environmental dynamics, deep models may capture intrinsic noises in CSI data, which decreases the performance. Moreover, in real-world scenarios, the training data requires labor-intensive data collection and annotations, which limits the capacity of normal deep learning models. In this

thesis, we seek to tackle these problems by proposing new learning architectures and adaptation methods.

2.2 Overview of Learning Algorithms for WiFi-based Sensing

As this thesis explores new learning algorithms for WiFi-based sensing, we review the recent progress of machine learning and deep learning. The connections and how to apply these algorithms to WiFi-based sensing are highlighted.

2.2.1 Traditional Statistical Learning

Statistical learning has two main purposes: classification and regression. For WiFi-based sensing, activity recognition and gesture recognition tasks belong to classification problems. The CSI data, which is a two-dimensional matrix, consists of temporal and spatial axes. For feature extraction, Principal Component Analysis (PCA) [39] analyzes the main components via Singular Value Decomposition (SVD). Then to obtain temporal patterns for CSI data, many statistics can be computed such as mean and variance. Such features can be classified by traditional statistical learning approaches such as Support Vector Machine (SVM), K-Nearest Neighbors (KNN), random forest and decision tree [40]. These approaches help recognize different category by distinguishing their statistical characteristics and training a fitting classifier. The drawback lies in the difficulty of extracting robust features from high-dimensional CSI data, which decreases the generalization ability of traditional statistical classifier.

2.2.2 Deep Representation Learning

To obtain robust representations on high-dimensional data such as images and CSI, deep learning comes into existence. Spatial features are extracted by Convolutional Neural Network (CNN) in an automatic manner [1]. Specifically, CNN is a specialized sort of artificial neural networks. It yields remarkable performance for many high-level tasks in computer vision such as object detection and segmentation [10].

In general, a feed forward network consists of multiple layers including convolution, pooling, activation and fully connected layers. Lecun et al. [1] firstly proposed CNN and LeNet-5 for hand-written digit recognition. The famous LeNet-5 is shown in Figure 2.1. Each layer of LeNet-5 is a function whose input is the output of its previous layer. Convolutional layers extract features by mapping a patch to a point by convolution operations of the patch and a kernel. Then pooling layers help subsample the feature map and thus reduce the dimension. Activation layers such as ReLU increase non-linearity of CNN and fully connected layers serve as the classifier. The optimization objective of CNN is a loss function. For a classification problem, we usually employ cross-entropy loss, which measures the KL-divergence of the predictions of CNN and the ground truth. The whole network is optimized by back-propagation in an automatic manner. For WiFi-sensing, 1D or 2D CNN can extract spatial and temporal features. The limitation is the size of the convolution kernel which only focuses on local patterns. If the length of patterns is long such as temporal data, then CNN may not be effective enough.

FIGURE 2.1: The architecture of LeNet-5. (Figure source: Lecun et al. [1])

To capture both long-term and short-term patterns, Recurrent Neural Network (RNN) was proposed [11]. Long Short-Term Memory (LSTM) [41] is a specialized design of RNN, which uses multiple gate units to forget or memorize specific signals. LSTM is suitable for representation learning of sequence data, which is widely used in natural language processing [42] and temporal data modeling [43]. However, LSTM requires sufficient training data and the computational complexity of its training is high when compared to CNN. For CSI data, when long-term patterns are important, LSTM is a panacea that learns to memorize these patterns from the temporal axis.

2.2.3 Few-shot Learning

A large scale of CSI samples are hard to collect, which requires many volunteers conducting specified actions. Therefore, when we adopt traditional deep neural networks based on supervised learning, it is quite labor-intensive and time-consuming to obtain the training data. Few-shot learning aims to tackle the problem of lack of annotated data [44]. Suppose that we only have several CSI gesture samples for training. Few-shot learning is able to learn effective features based on more powerful metric-based or similarity-based losses. This thesis explores this direction for few-shot gesture recognition [45] that employs a Siamese neural network for efficient learning. Few-shot learning techniques enable WiFi-based sensing with limited data and annotations.

2.2.4 Transfer Learning for Robust Feature Extraction

When the CSI data is collected, the surrounding environmental information is also captured, which causes that the model trained in one place cannot generalize well in a new environment. In transfer learning, fine-tuning and domain adaptation are two solutions to this problem.

- **Pretraining and fine-tuning.** If we aim to train a robust CSI feature extractor that is suitable for most circumstances, a large scale of CSI samples are required. For example, in computer vision field, one essential factor is the collection of ImageNet that includes 100 million labeled images for 1000 visual categories. If a model is aimed at visual recognition, pre-training it on ImageNet [46] becomes extremely necessary. In WiFi-based sensing, the scales of CSI data for most applications are small for deep learning model. Though new large-scale datasets such as Widar [47] were proposed recently, a general network architecture with pre-trained parameters for CSI-based sensing is still lacking.
- **Domain adaptation.** The difference caused by environments or layouts is a form of domain shift in the learning theory of domain adaptation [48]. Domain adaptation algorithms help transfer knowledge from the original environment (*i.e.* source domain) to a new environment (*i.e.* target domain).

By minimizing statistical measures of domain shift, models can be adapted to the new environment in an unsupervised manner [49]. In this thesis, novel domain adaptation algorithms are proposed to empower stable and comprehensive domain alignment, and we will review domain adaptation from several perspectives in the next section.

2.2.5 Cross-Modal Learning for Annotation Assistance

In smart sensing field, images and wireless signals are two kinds of modalities. They have their own characteristics that decide their functionalities. Images have high granularity and are suitable for high-level tasks such as visual tracking and semantic segmentation. However, they suffer from the privacy issue in smart homes, which is not a problem for WiFi-based sensing. Different from images that can be directly understood and then annotated in pixel level, WiFi CSI data can only be annotated with simple labels that refer to activity or gesture type. To empower high-level tasks such as skeleton extraction using wireless signals, cross-modality supervision aims to leverage visual ground truth for training wireless signals, which empowers pose estimation based on wireless signals [50].

2.2.6 Multi-Modality Learning for Comprehensive Sensing

If a camera-based method is integrated with wireless signals that can detect activities through wall, two modalities learn from each other's strength and close the sensing gap together. Deep multi-modal learning contributes to more sensible ways of combining and connecting two kinds of features. For example, Zou et al. integrate both visual and CSI data for accurate multi-modality activity recognition [51].

2.2.7 Model Compression for Edge Computing

Another challenge for WiFi-based sensing is how to deploy an algorithm in edge devices such as WiFi routers. Smart routers have some computational resources but cannot support real-time inference of deep networks. To accelerate the speed and decrease the computations, model compression comes into existence. It mainly

includes three ways including network pruning, quantization and distillation. Network pruning aims to reduce the number of parameters by pruning some connections of neurons or CNN kernels [52]. Quantization methods quantize all the weights to integers. Knowledge distillation transfers knowledge from a deep network to a shallow network, which is explored for CSI-based sensing in this thesis [53].

2.3 Domain Adaptation Algorithms

A major assumption in many machine learning algorithms is that the training and future testing data share the same feature space and thus have the same distribution. However, in many real-world scenarios, this assumption does not hold such as the temporal shift of WiFi-based indoor localization [54]. To tackle this problem, domain adaptation aims to transfer knowledge from a label-rich source domain to a label-scarce or even unlabeled target domain [49]. Denote $\mathbb{P}$ the source distribution and $\mathbb{Q}$ the target distribution. The problem is that $p(x) \neq q(x)$. For the same task, the assumption of domain adaptation is that the conditional output distribution is invariant, *i.e.* $p(y|x) = q(y|x)$, where x, y are the input data and output labels, respectively [12]. As such, most domain adaptation algorithms deal with the divergence of distributions between the training and future testing data for the same task.

In deep learning, domain adaptation tackles two important issues: lack of data in the testing scenario and the dataset bias between the training and testing data. As discussed in Section 2.2.4, though we can train deep models using existing training dataset, there exists domain shift between the testing data and the training data. To deal with it, collecting data in the new environment (*i.e.* testing scenario) is rather expensive and even unavailable in some situations. Domain adaptation helps adapt the model to the new environment using labeled data from the source domain and the label-scarce or unlabeled target domain. Using a few labels in the target domain is named Supervised Domain Adaptation (SDA) and using no label in the target domain is called Unsupervised Domain Adaptation (UDA). This thesis mainly focuses on UDA. Hence, we begin with the learning theory of UDA, and review the recent UDA methods from the perspectives of alignment approach, characteristics of learning representations and type of domain shift. The current domain adaptation benchmarks are also introduced.

2.3.1 Learning Theory of Domain Adaptation

Most deep domain adaptation methods aim to minimize the source and target domain discrepancies, which is motivated by Ben-David et al. [48] which derives an upper bound for the expected error of the target samples $\epsilon_T(h)$ in the learning theory of domain adaptation.

Theorem 2.3.1. For a hypothesis $h \in \mathcal{H}$, the expected error of the target domain is bounded by

$$\epsilon_T(h) \leq \epsilon_S(h) + \frac{1}{2} \hat{d}_{\mathcal{H} \Delta \mathcal{H}}(\mathcal{U}_S, \mathcal{U}_T) + \lambda + C_0, \quad (2.1)$$

where $\epsilon_S(h)$ is the source error, $\hat{d}_{\mathcal{H} \Delta \mathcal{H}}(\mathcal{U}_S, \mathcal{U}_T)$ is the empirical estimation of the $\mathcal{H}$ -divergence on samples $\mathcal{U}_S, \mathcal{U}_T$, λ is the risk of an ideal hypothesis for both domains and C_0 is a constant depending on the complexity of hypothesis space (e.g. VC dimension, Rademacher complexity or metric entropy).

In Theorem 2.3.1, the expected error of the target domain is bounded by the source error, the discrepancy term, the risk of the ideal hypothesis and the complexity of hypothesis space. The source error is fixed since the labels are available in the source domain. For a specific model, the complexity of hypothesis space is fixed. Then the discrepancy between two domains and the ideal risk are the main factors of the expected error of the target domain. As the ideal risk is thought to be small, most UDA approaches propose to minimize the discrepancy. Though in this thesis we discuss that the ideal risk may not be neglected, here we mainly introduce how UDA minimizes the discrepancy.

To minimize the distribution discrepancy, recent deep domain adaptation approaches are categorized to two main streams: statistics-based or adversarial-based approaches.

2.3.2 Domain Adaptation by Global Statistics

Statistics-based methods propose to measure the discrepancy by statistics and then minimize it. To this end, many statistics for moment matching are employed, in which Maximum Mean Discrepancy (MMD) plays an essential role. MMD [55] has been proved to be effective in the *two-sample* testing. Given samples following

probability distributions p and q for the source and target domain respectively, acceptance or rejection decisions are made for a null hypothesis $p = q$ in the *two-sample* testing. Pan et al. pioneer transfer component analysis that learns some transfer components across domains in a Reproducing Kernel Hilbert Space (RKHS) using MMD [56]. Multiple kernel variant of MMD (MK-MMD) is then proposed by Gretton et al. [57] for better stability and stronger two-sample test power.

Assume $\mathcal{H}_k$ is the reproducing kernel Hilbert space (RKHS) endowed with a kernel $k(x, x')$. Its element $k(x, \cdot)$ satisfies the reproducing property which is associated with the implicit feature map $\phi(x)$ where $k(x, x') = \langle \phi(x), \phi(x') \rangle_{\mathcal{H}_k}$. The kernel embedding $\mu_{\mathbf{x}}(p)$ stands for a probability distribution p by an element in RKHS $\mathcal{H}_k$. Moreover, the expectation of any RKHS function $\mathcal{F}$ can be evaluated as an inner product in $\mathcal{H}_k$ using the mean embedding $\bar{\mu}_{\mathbf{x}}$, i.e. $\mathbb{E}_{\mathbf{x}}[\mathcal{F}(x)] \triangleq \langle \bar{\mu}_{\mathbf{x}}, \mathcal{F} \rangle$. The MMD between the source and target distributions p and q is defined as the RKHS distance $d_k(p, q)$ between their mean embeddings. Its squared formulation is defined as

$$d_k^2(p, q) \triangleq \|\mathbb{E}_p[\phi(x_s)] - \mathbb{E}_q[\phi(x_t)]\|_{\mathcal{H}_k}^2. \quad (2.2)$$

The MK-MMD holds the property that $p = q$ iff $d_k^2(p, q) = 0$. To select a kernel $k(x_s, x_t) = \langle \phi(x_s), \phi(x_t) \rangle$ for the associated feature map ϕ , we define the convex combination of m_k positive definite functions kernels k_u ,

$$\mathcal{K} \triangleq \left\{ k = \sum_{u=1}^{m_k} \beta_u k_u : \sum_{u=1}^{m_k} \beta_u = 1, \beta_u \geq 0, \forall u \right\}. \quad (2.3)$$

As illustrated in Gretton et al. [57], the multi-kernel choice can ensure a stronger test power and low test error, resulting in a superior method for optimal kernel selection. Hence, the multi-kernel set $\mathcal{K}$ used in MK-MMD can intensify the two-sample test power, which is more robust than MMD. For deep models, MMD or MK-MMD can be simply appended as a regularization term that calculates the discrepancy measure of deep features between two domains [58]. Long *et al.* further introduce an optimal multi-kernel selection method for mean embedding matching [3, 59].

To connect MK-MMD with the learning theory of UDA, Long et al. [3] employ a Parzen window classifier [60] and derive the following result:

Theorem 2.3.2. The $\mathcal{H}$ -divergence between two domains can be bounded by the domain discrepancy measurement as follows:

$$\hat{d}_{\mathcal{H}}(\mathcal{U}_{\mathcal{S}}, \mathcal{U}_{\mathcal{T}}) \leq 2(1 + d_k(\mathcal{U}_{\mathcal{S}}, \mathcal{U}_{\mathcal{T}})) + C_1 \quad (2.4)$$

where $d_k(\mathcal{U}_{\mathcal{S}}, \mathcal{U}_{\mathcal{T}})$ is the discrepancy measurement (*i.e.* MK-MMD) and C_1 is a constant depending on the complexity of hypothesis space.

It is shown that minimizing the measures of discrepancy help tighten the upper bound of the discrepancy of two domains, which further contributes to minimization of the expected error in the target domain. Apart from MMD, more statistical measures have been developed. Sun et al. [61] propose to measure the shift by covariance matrices. Statistically invariant embedding is proposed to measure the discrepancy on a Riemannian manifold [62]. Statistics-based UDA methods have either closed-form solution [56] or stable training procedure [3], which is very effective for the WiFi-based indoor localization problem and cross-domain object recognition.

2.3.3 Adversarial Domain Adaptation

Inspired by Generative Adversarial Networks (GAN) [63], Adversarial Domain Adaptation (ADA) designs a two-player adversarial game between the feature extractor and an auxiliary domain classifier. The domain classifier learns to distinguish the source features from the target features while the feature extractor learns to deceive. Ganin and Lempitsky [2] firstly proposed to learn domain-invariant features by Domain-Adversarial Training (DAT) using a Gradient-Reverse Layer (GRL), which has inspired many UDA methods thereafter. For example, Zhang et al. [64] propose a new divergence for distribution comparison based on minimax optimization and Wang et al. [65] discover that filtering unrelated source samples helps avoid negative transfer in DAT. Adversarial Discriminative Domain Adaptation (ADDA) tries to fool the label classifier by adversarial training but not in an end-to-end manner. CyCADA [66] and PixelDA [67] leverage GAN to conduct both feature-level and pixel-level domain adaptation, which yields significant improvement yet the network complexity is high.

Theoretically, ADA can be explained by the learning theory of UDA. Let us elaborate DAT with GRL as an example. Consider the $\mathcal{H}\Delta\mathcal{H}$ -distance in Theorem 2.3.1:

$$d_{\mathcal{H}\Delta\mathcal{H}}(\hat{\mathcal{D}}_S, \hat{\mathcal{D}}_T) = 2 \sup_{h_1, h_2 \in \mathcal{H}} \left| \Pr_{\mathbf{z} \sim \hat{\mathcal{D}}_S} [h_1(\mathbf{z}) \neq h_2(\mathbf{z})] - \Pr_{\mathbf{z} \sim \hat{\mathcal{D}}_T} [h_1(\mathbf{z}) \neq h_2(\mathbf{z})] \right|, \quad (2.5)$$

where $\hat{\mathcal{D}}_S$ and $\hat{\mathcal{D}}_T$ denote the distributions of the source and target feature space $\mathcal{Z}_S$ and $\mathcal{Z}_T$, respectively. Given the fixed feature extractor G_e , $\hat{\mathcal{D}}_S$ and $\hat{\mathcal{D}}_T$, we consider a family of label predictors $\mathcal{H}_p$. Inspired by [68], we can simply assume a family of domain classifiers $\mathcal{H}_d$ to be rich enough to contain the symmetric difference hypothesis set of $\mathcal{H}_p$, such that $\mathcal{H}_p \Delta \mathcal{H}_p = \{h | h = h_1 \oplus h_2, h_1, h_2 \in \mathcal{H}_p\}$ where $\oplus$ is the XOR-function. Then we can easily demonstrate that the optimal domain classifier gives an upper bound for the $\mathcal{H}_p \Delta \mathcal{H}_p$ -distance:

$$\begin{aligned} d_{\mathcal{H}_p \Delta \mathcal{H}_p}(\hat{\mathcal{D}}_S, \hat{\mathcal{D}}_T) &= 2 \sup_{h \in \mathcal{H}_p \Delta \mathcal{H}_p} \left| \Pr_{\mathbf{z} \sim \hat{\mathcal{D}}_S} [h(\mathbf{z}) = 1] - \Pr_{\mathbf{z} \sim \hat{\mathcal{D}}_T} [h(\mathbf{z}) = 1] \right| \\ &\leq 2 \sup_{h \in \mathcal{H}_d} \left| \Pr_{\mathbf{z} \sim \hat{\mathcal{D}}_S} [h(\mathbf{z}) = 0] + \Pr_{\mathbf{z} \sim \hat{\mathcal{D}}_T} [h(\mathbf{z}) = 1] - 1 \right|. \end{aligned} \quad (2.6)$$

The domain classifier in ADA is $h(\mathbf{z})$ in Eq.(2.6). Therefore, when we optimize the domain classifier $h(\mathbf{z})$, the upper bound of the $\mathcal{H}_p \Delta \mathcal{H}_p$ -distance is tightened, which further tightens the expected error of the target domain.

Generally speaking, ADA is attractive due to excellent alignment performance and the flexibility of the plug-and-play GRL. Nevertheless, adversarial training suffers from training instability. Since the convergence cannot be guaranteed for a deep adversarial network, ADA usually requires manual hyper-parameters tuning to achieve the convergence, which hinders the real-world usage of ADA. In many applications, domain adaptation is expected to work in an automatic manner. This issue motivates the work in Chapter 6.

2.3.4 Transferability and Discriminability

From the perspective of feature representations, we can categorize UDA approaches to two purposes: enhancing the transferability or discriminability. In the common feature space shared by two domains, if there are more common feature subspaces, the transferability is better. The discriminability denotes the level of how these

subspaces contribute to representing the two domains. Both of them are very important for domain adaptation and they reflect different characteristics of features. On the one hand, some UDA methods that only improve the transferability may not pay attention to the decreasing discriminability. On the other hand, strong discriminability in the source domain may hinder the transferability. Hence, a trade-off should be achieved.

Transferability in UDA. Most UDA methods aim to enhance the transferability of feature representations. According to the learning theory of UDA [48], this could be achieved by minimizing the distribution discrepancy between the source and the target domain in the feature space. The statistics-based methods [58, 69] and ADA [2, 70] can achieve it for deep learning models. The transferability of features can be enhanced by pixel-level adaptation [67] and [66].

Discriminability in UDA. Though ADA methods contribute to learning domain-invariant features, they could hinder the discriminability of feature representations in the target domain [71]. To enhance the discriminability, semantic features attract more attention for domain adaptation. Many approaches are proposed to learn discriminative features by the boundary adjustment [72, 73]. MSTN employs semantic features with self-training for domain adaptation [74]. BSP penalizes the overlarge transferability for a better trade-off [71]. These methods either amend the ADA by adding extra regularizers or build self-training algorithms according to their feature embeddings without consideration of domain-invariant feature learning. To overcome these issues, this thesis proposes a simple yet powerful adversarial approach to simultaneously minimize the domain discrepancy and inherit good discriminability from the source domain in Chapter 7.

2.3.5 Covariate Shift and Label Shift

Domain Adaptation with Covariate Shift. Standard domain adaptation addresses the problem of transferring knowledge from a source domain to a target domain with covariate shift [12]. Assuming $p(y|x) = q(y|x)$, such covariate shift can be minimized by aligning the marginal distribution, *i.e.* $p(x) = q(x)$, since the learning theory of domain adaptation indicates that the target error is bounded by the source error and the cross-domain distribution discrepancy [48]. As discussed

before, covariate shift can be handled by domain alignment methods such as Maximum Mean Discrepancy (MMD) [69] and ADA [75]. Apart from domain alignment, self-training and prototype-based methods iteratively select high-confidence target samples for training [76–78]. Similarity-based UDA learns a metric that can obtain semantic representations [79, 80]. Though these methods significantly improve the performance of UDA, they may result in negative transfer under the label shift.

Domain Adaptation with Label Shift. Even when two domains share the same conditional data distributions, *i.e.* $p(x|y) = q(x|y)$, different label distributions often occur, *i.e.* $p(y) \neq q(y)$. One such situation is the well-known long-tailed data distribution that describes the nature of visual categories [81]. For example, in face recognition, the range loss was designed to handle the long-tailed problem by reducing overall intra-class variations [82]. As to predicting the label shift, Lipton *et al.* propose a distribution estimator to detect it [83]. Then the first error bound for label-shift problem without any labels in the target domain is derived [84]. *Note that covariate shift is also considered in IDA while these approaches only study label shift, which makes the IDA problem more challenging.* Pan *et al.* proposed weighted MMD (WMMD) [85] for biased cross-domain datasets. Recently, open set domain adaptation [86] and partial domain adaptation [87] were proposed on the basis of whether the cross-domain label space is overlapped. Due to the lack of some categories in the source domain, these two scenarios also belong to special Imbalanced Domain Adaptation (IDA) cases, which will be introduced in Chapter 8.

2.3.6 Other Domain Adaptation Categories

Unsupervised feature learning. Another line of approaches employ unsupervised learning for UDA such as autoencoder, which enables unsupervised image-to-image translation by learning pixel-level features [88]. Unsupervised learning helps extract features from the target domain and benefits the search of a common feature space. Ghifary *et al.* [89] employed a decoder network for pixel-level adaptation, and Domain Separate Network (DSN) [90] further leveraged multiple decoder networks to learn domain-specific features. These approaches treat the decoder network as an independent component for augmented feature learning that is irrelevant to domain alignment [91].

Adjusting decision boundary. Though domain alignment based approaches are quite effective, they usually neglect the problem of the decision boundary while aligning the features. Shu et al. [72] adopted virtual adversarial training to address the drawbacks of merely domain alignment approaches, which has gained a huge success. Other methods were developed to deal with the boundary problem in domain alignment such as Co-DA [92] and MCD [73], achieving simultaneous domain and class alignment. In Chapter 6, our proposed method is a new generic approach for invariant feature learning, which can be easily integrated with these existing UDA methods to provide a better domain alignment performance.

2.3.7 Domain Adaptation Benchmarks

To evaluate domain adaptation methods, many datasets were collected, and here we introduce the prevailing benchmarks on visual and WiFi domain adaptation.

Digits [93]. There are five digits datasets **MNIST**, **MNIST-M**, **USPS**, **SVHN** and **Synthetic Digits** (SYN-DIGIT)), all of which consist of 32×32 images. Five transfer tasks with multiple levels of domain difference are formed as: **MNIST**→**USPS**, **USPS**→**MNIST**, **SVHN**→**MNIST**, **SYN-DIGIT**→**SVHN**, **MNIST**→**MNIST-M**. The protocol to evaluate on **Digits** is settings of **DANN** [2] that uses the official training splits in two domains for training and evaluates the model on the testing split.

Office-31 [94] is the most commonly used dataset for visual domain adaptation. It consists of 31 categories and totally 4652 images collected from three domains: *Amazon* (**A**) with 2817 images, *Webcam* (**W**) with 795 images and *DSLR* (**D**) with 498 images that are obtained from Amazon¹, web camera and digital SLR camera, respectively. The **W** and **D** domains are quite similar but the **A** domain has more images from distinct sources. The sample size of *Office-31* is quite small and the image numbers of three domains are imbalanced. The evaluation is conducted across all six transfer tasks: **A**→**W**, **D**→**W**, **W**→**D**, **A**→**D**, **D**→**A** and **W**→**A**.

ImageCLEF-DA² is a public benchmark for the ImageCLEF 2014 domain adaptation challenge. The organizers selected 12 common categories from three public datasets to form three domains: *Caltech-256* (**C**), *ImageNet ILSVRC 2012* (**I**)

¹<https://www.amazon.com/>

²<http://imageclef.org/2014/adaptation>

FIGURE 2.2: The example images of Office-Home.

and *Pascal VOC 2012* (**P**). This dataset is more structured with 50 images in each category and 600 images in each domain. Six transfer tasks are formed as: $\mathbf{I} \rightarrow \mathbf{P}$, $\mathbf{P} \rightarrow \mathbf{I}$, $\mathbf{I} \rightarrow \mathbf{C}$, $\mathbf{C} \rightarrow \mathbf{I}$, $\mathbf{C} \rightarrow \mathbf{P}$ and $\mathbf{P} \rightarrow \mathbf{C}$.

Office-Home³ [95] is a very challenging dataset comprising around 15500 images from 4 significantly different domains: *Artistic images* (**Ar**), *Clip Art* (**Cl**), *Product images* (**Pr**) and *Real-World images* (**Rw**). Moreover, each domain contains 65 categories of everyday objects in office and home. The adaptation on *Office-Home* is very difficult due to the large category number and enormous domain shift. Some examples of four domains are shown in Fig. 2.2.

VisDA-2017 [96] is a very challenging domain adaptation task which contains 280K images in 12 classes. **Synthetic** images are 3D models generated by computers and **Real** domain are collected from real-image datasets.

For **Image-CLEF**, **Office-Home** and **VisDA-2017**, the full training protocol is developed in [3] that employs all images from the source domain and the target domain for training.

WiFi Gesture Recognition [97] consists of six gestures recorded by Channel State Information (CSI) [98]. Each CSI sample is a 2D matrix that depicts the gesture with the surrounding layout environment. The dataset is collected using the platform in this thesis. The CSI data collected in two environments forms two domains, which formulates a spatial adaptation problem. The evaluation protocol is the same as the protocol used in *Digits*.

³<http://hemanthdv.org/OfficeHome-Dataset>

2.4 Conclusion

In this chapter, we have surveyed some major sensing technologies, learning techniques for WiFi-based sensing and domain adaptation algorithms. For sensing techniques, Camera, IR, IMU, RF, lidar and WiFi have been reviewed. As WiFi-based sensing constitutes one of the main contributions of the thesis, the learning techniques for WiFi-based sensing have been reviewed. The limitations of WiFi-based sensing applications include the labor-intensive annotations of data, being vulnerable to spatial and temporal dynamics, lack of pixel-level annotations and lack of robust feature extractor. These limitations can be partially tackled by recent learning techniques, which motivates second part of this thesis. The environmental dynamics are defined as domain shift in statistical learning, which is a common problem in various sensing technologies such as style difference in visual sensing and spatial shift in WiFi-based sensing. To deal with it, domain adaptation algorithms have been reviewed. Though existing UDA methods yield satisfactory performance, some of them suffer from convergence issue, lack of pixel-level adaptation and the difficulty of dealing with imbalanced classes. These limitations have been pointed out in this chapter, which motivates the domain adaptation research in this thesis.

Part I

Smart Sensing

Chapter 3

WiFi-based IoT-enabled Sensing Platform

In this chapter, we propose a new WiFi-based IoT-enabled sensing platform. The motivation and introduction are provided in Section 3.1. Section 3.2 gives a brief introduction to CSI data and connects CSI with human motions. Section 3.3 precisely describes the framework of our IoT-enabled sensing platform. In Section 3.4 and Section 3.5, we present two case studies for occupancy detection and sedentary behavior monitoring based on our platform and machine learning techniques. Then the experiments and evaluations are conducted in terms of accuracy, robustness and scalability, as illustrated in Section 3.6. At last, we conclude this chapter in Section 3.8.

3.1 Introduction

Occupancy information plays an essential role in smart buildings and homes in terms of energy consumption estimation, human behavior analysis and indoor security monitoring. With the rapid development of the IoT, this information can be obtained by the cognitive computing technologies, which are introduced into smart homes by the integration of sensors, actuators, controllers and artificial intelligence. Various applications are enabled, of which the occupancy sensing is a

significant one [99]. It refers to estimating the occupancy information and simulating the situation of indoor environments by means of the analysis of data from kinds of sensor networks deployed in smart buildings.

With occupancy sensing using IoT devices, the systems in smart buildings or homes are able to actively obtain the information of occupancy, such as number of occupants, activities and even gestures. For instance, occupancy detection and human identification help detect any intruders automatically. In smart homes, the system can adjust the light and temperature according to the occupancy, which conserves the energy. For children and the elderly, it can monitor their activities such as falling to prevent potential hazard. Generally speaking, IoT-based occupancy sensing provides an active detected input for the smart control systems, so that they can make correct judgment.

The conventional approaches to perceive the occupants mainly resort to kinds of sensors, such as ultrasound, CO_2 and thermal sensors [23]. These works focus on coarse sensing including occupancy detection or counting [100], but as the machine learning techniques fortify the development and applications of artificial intelligence, more fine-grained sensing comes in existence. Researchers utilize sensors or wearable devices to achieve remarkable results [15]. However, the challenge in fine-grained sensing lies in finding a trade-off between performance and efficiency. For instance, it is crucial to leverage a device-free manner due to convenience. In general, most prevailing solutions rely on dedicated sensors such as cameras [101], RF tags [102] and accelerometers [5]. These solutions require either toilsome installations or use of wearable devices. The camera-based methods provide highest granularity of activity recognition via computer vision algorithms, but they are highly impacted by illumination and occlusion. More importantly, due to privacy concerns, it is usually not so acceptable in smart homes. Another good solution applies accelerometer equipped in most smart phones. Nevertheless, people do not necessarily carry cellphones or particular devices all the time under indoor circumstances. Radio frequency based approaches show their advantages in low consumption and device-free manner. However, this kind of system requires many RF links in a single room, and the installation cost decreases its flexibility and availability. Some other solutions such as location-based methods need auxiliary information so that they cannot recognize precise activity happening in the same place.

Recently, it is observed that the motions of occupants impact the WiFi signals to some extent. Based on this insight, researchers manage to use Received Signal Strength Indicator (RSSI) [103] and Channel State Information (CSI) [24] to depict the human activity [104, 105]. WiFi-based technology has three advantages compared with other methods. Firstly, it is tailored for practical deployment as a ubiquitous system. As WiFi technique has been developed rapidly, the indoor coverage of wireless networks has become more and more widespread. Thus, we do not need to install specific sensors or RF links. Secondly, WiFi-based approaches employ the pattern extracting of wireless signals in the physical layer, and thus do not raise issues of privacy. Thirdly, it is able to provide relatively accurate sensing with minimal infrastructure requirements, which is a sensible trade-off between availability and performance.

When WiFi-based methods are applied to practical use, there still remain several challenges such as scalability of platforms, and lack of resolutions of wireless signals, etc. In this chapter, we address these challenging problems and present a novel technology that utilizes commercial off-the-shelf (COTS) Wi-Fi enabled IoT devices for occupancy sensing in smart homes. Our system employs the fine-grained CSI data from the physical layer in the wireless connection between the transmitter (Tx) and receiver (Rx). It is designed for occupancy sensing with high scalability and performance, and it is composed of three parts, including an innovative IoT platform for sensing, cloud server for computing and user end for notification. We firstly upgrade OpenWrt system and enable it to report exquisite CSI on the IoT devices. Then according to different application, the data can be processed in either a coarse or fine-grained way. To demonstrate the performance of our system, we then develop and realize two challenging applications, referring to occupancy detection and sedentary behavior monitoring by human activity recognition. For applications of coarse sensing such as occupancy detection, we utilize CSI curve similarity to achieve the goal. It is completely implemented on the IoT devices. For applications of fine-grained sensing for sedentary behavior monitoring, we classify human motions to static or dynamic activities and then employ statistical learning to conduct fine-grained classification. The evaluation in typical indoor environments shows that two applications achieve sensible performance in real time.

Contributions: We summarize the main contributions as follows:

- An agile WiFi-enabled IoT system for occupancy sensing is proposed. To the best of our knowledge, it is the first platform which completely integrates CSI-based WiFi devices, cloud server and user end. It makes a good trade-off between performance and availability.
- The OpenWrt operating system is upgraded with CSI reporting function for COTS routers. The fine-grained CSI data and scalable platforms provide a good opportunity to promote existing applications of WiFi-based systems.
- To demonstrate our framework, we elaborate a coarse sensing case, occupancy detection, based on similarity measurement of CSI curves, and realize it on the IoT device. As far as we know, it is the first WiFi-based application purely implemented on the router.
- For fine-grained sensing, we investigate the characteristics of wireless signals influenced by sedentary behavior and propose a kernel density estimation based method to conveniently distinguish dynamic in-place activities and still postures during sedentary period. Then, we creatively utilize both frequency diversity and temporal information concealed in CSI to design our system namely CareFi, which is composed of four innovative phases.
- We evaluate both the detection and recognition systems in terms of accuracy, robustness and scalability. The experimental results show the superiority of our platform.

3.2 Preliminary: Channel State Information

In wireless communication, channel state information is a representation of the channel properties of a communication link. During the propagation of wireless signals, they are affected by physical environment, which leads to reflections, diffractions and scattering [106]. These phenomena can be described by CSI. Modern WiFi devices adopt Orthogonal Frequency Division Multiplexing (OFDM) at the physical layer and obey IEEE 802.11n/ac standard that permits multiple transmit and receive antennae for multiple input, multiple output (MIMO) communication. Therefore, CSI reveals fine-grained characteristics of wireless signals combining effect of time delay, amplitude attenuation and phase shift of multiple paths on each

communication subcarrier. Since RSSI is only a superimposition of multi-path signals, CSI naturally owns higher resolution. The WiFi signals can be modeled as Channel Impulse Response (CIR) $h(\tau)$ in frequency domain:

$$h(\tau) = \sum_{l=1}^L \alpha_l e^{j\phi_l} \delta(\tau - \tau_l), \quad (3.1)$$

where α_l and ϕ_l represent the amplitude and phase of the l -th multi-path component respectively, τ_l is the time delay, L indicates the total number of multi-path and $\delta(\tau)$ denotes the Dirac delta function. However, only clusters of multipath components are differentiable because of limited WiFi bandwidth. In this case, the OFDM receiver provides a sampled version of the signal spectrum at subcarrier level, which comprises amplitude attenuation and phase shift via complex number. These estimation can be represented by:

$$H_i = ||H_i|| e^{j\angle H_i} \quad (3.2)$$

where $||H_i||$ and $\angle H_i$ are the amplitude and phase of i -th subcarrier respectively.

Theoretically, CSI phase information is supposed to be more robust with less variations. Nevertheless, it is opposite due to hardware imperfections and environmental variations [107]. The carrier frequency of a device can drift by up to 100kHz for 5GHz band. In practical applications, it brings about extreme complexity to accomplish calibration and denoising. Thus, we employ amplitude information to design our system.

3.3 Agile WiFi-enabled IoT Platform for WiFi-based Sensing

In this section, we introduce our agile WiFi-enabled IoT platform for occupancy sensing, which consists of an IoT platform for WiFi device, cloud server for complex computations and user terminal for display of results. In the platform design, it is significant to ensure the real time with the consideration of network and computing overload. For example, if CSI measurements are transmitted to the server at a high sampling rate, the successive transmission will put a great deal of burden on the

FIGURE 3.1: Overview of the system architecture.

network of the server. Meanwhile, if we process every frame of CSI in the server, it may cause computing resource starvations. After initial experiments and investigations, we design our real-time occupancy sensing system shown in Figure 3.1, integrating embedded operating system OpenWrt, lightweight messaging prototype MQTT and cloud server. The IoT platform refers to two devices supporting WiFi communication, which is installed our innovative OpenWrt firmware to enable the capability of reporting CSI from the physical layer directly. To reduce the pressure on network transmission, we conduct some reasonable calculations on the device according to various applications, and these results are transmitted to the server for further processing. The cloud server is responsible for machine learning framework and message distribution as *MQTT broker*. The sensing results are finally displayed on the user terminal, such as laptops and mobile phones.

3.3.1 IoT Platform

The IoT platform takes charge of real-time collection of CSI and primary computations, which enables rapid and scalable deployment. Conventional CSI tools used in almost all researches of CSI based sensing, such as Intel 5300 NIC tool [24] or Atheros 9390 tool [108], consist of a commodity router as transmitter and

a mini PC or specific devices with WiFi NIC card as receiver. Generally speaking, they are sufficient for research but not for practical applications. There exist two limitations in terms of these platforms. Firstly, the receiver has to be laptops with specific network card, which restrains the practicability and scalability. For instance, Wang et al. [109] employ more than 10 laptops as receivers in the implementation of an indoor localization system. The situation leads to high cost and inconvenience of maintenance, impeding the realization of large-scale systems. The second limitation lies in the resolution of CSI measurements. Intel 5300 NIC tool [24] only provides CSI for 30 out of 56 subcarriers with 20Mhz bandwidth for each TX-RX pair. The CSI measurements are sampled which irreparably loses much necessary information in the physical layer.

To deal with these bottlenecks, we upgrade the Atheros CSI Tool [25] and develop a new firmware for OpenWrt system, which can be deployed in many COTS routers. Our platform only requires two devices with Atheros WiFi module, such as laptop, router and development board, which boosts the convenience and reduces the cost substantially. The CSI can be directly reported in the OpenWrt system. We choose OpenWrt since it supports a great number of routers and boards. It is so lightweight that much redundancy of normal operating systems is removed. Equipped with the new firmware, the IoT devices have the ability to report fine-grained CSI with complete 114 subcarriers with 40Mhz bandwidth under 5Ghz. Moreover, we can directly conduct some essential computations on the IoT devices.

We aim at building a generic platform where kinds of applications can be developed without involving lots of efforts. Thus, we classify the tasks of occupancy sensing in smart homes into two categories, coarse and fine-grained sensing. To make full use of the computing resources on IoT devices and cloud server, we render our platform in two modes. For coarse sensing tasks including room-level localization and occupancy detection, the computational complexity is tolerable for IoT devices. We manage to implement the algorithms such as local classifier on the board to complete the task. Then the device directly transmits the result to the server and let the users obtain real-time information without using the computing resources of the server. Fine-grained sensing tasks mainly refer to activity recognition and even gesture recognition. These works require advanced signal processing and machine learning techniques to achieve better performance, and these cost enormous computing resources. Furthermore, real-time transmission of CSI

measurements can bring about network overload. To tackle these challenges, we propose our recognition approach in Section 3.5, which extracts features from selected CSI data for transmission. Then the classification is conducted in the cloud server. This approach also works for other state-of-the-art methods, which mostly extract statistical features (such as mean, minimum, entropy) on the device and then transmit them.

3.3.2 Cloud Server and User Terminal

With the rapid development of cloud computing, the cloud server has huge advantages in workload resiliency and stability. In our system, we employ the cloud server, which presides over machine learning computations as well as message forwarding. In order to optimize network communication between lightweight OpenWrt and the server, MQ Telemetry Transport (MQTT) is introduced in our system. MQTT is an extremely simple lightweight messaging protocol, which is designed specially for constrained devices and low-bandwidth or high-latency networks. It has been successfully deployed in satellite links, medical and miniaturized devices [110]. With the help of MQTT, we manage to overcome the network overload.

In MQTT, the message is transmitted via TCP/IP protocol and it provides bi-directional connections in order to avoid prejudice. The MQTT is integrated in our system as follows:

- There are three roles in MQTT, including *publish*, *broker* and *subscribe*. The cloud server, IoT platform and user terminal play the role of *broker*, *publish* and *subscribe* respectively. The cloud server (*broker*) forwards the specific message to the user (*subscribe*) when the IoT platform (*publish*) pushes information to the cloud server. Our IoT platform (*publish*) generates sensing results and transmits the results to cloud server. Then the users (*subscribe*) are able to obtain required message. In our test, MQTT ensures the real-time notification in terms of IOS, Android and Web terminals.
- For MQTT *broker*, it offers quality of service (QoS) regarding different clients. In our system, we enable the highest quality of service for transmission.

- MQTT messages are delivered via matching *topic name*, which is a tag in each message. The IoT and subscribed devices of same user share same *topic name*. By this way, we make it possible to deploy our system in large scale of users.

After the MQTT pushes the message to the user terminal, the user interface should display real-time sensing results and be compatible with multisystems. In light of this, we develop the user interface based on Web, which supports most smartphones and laptops.

Based on our innovative platform, we propose two novel approaches to realize real-time applications. The occupancy detection is regarded as coarse sensing case, and it is purely implemented on the IoT device. In comparison, human activity recognition is a fine-grained occupancy sensing technology, and its machine learning framework is running on the server.

3.4 Case Study: Coarse Sensing for Occupancy Detection

Precise detection of occupancy is indispensable to improve the efficiency of building management system (BMS) [111]. It can be utilized for intrusion detection in a commercial area, thermal comfort adjustment in a smart building, and elderly monitoring at home. Traditional methods mostly rely on light, temperature and CO_2 measurements [23]. Based on our IoT platform, we propose an occupancy detection application as a case study of coarse sensing.

3.4.1 Insights: How Occupancy Affects Wireless Signals

Due to multi-path effects, human motion causes the variations of wireless signals that are characterized by CSI amplitude. To link the variation to common human activities, we conduct a preliminary experiment. We deploy two TP-Link N750 routers, each with one antenna, as transmitter (TX) and receiver (RX). The TX generates and sends a packet every 2 milliseconds and the same frequency of reception in RX. The distance between them is 1.5 meters, and the number of subcarriers

FIGURE 3.2: Curves of CSI subcarrier values from unoccupied (U) to occupied (O) scenarios. We can see that there exists obvious difference of curves between O and U.

FIGURE 3.3: STI values of one TX-RX pair. The occupants appear for four times indicated by four peaks.

is 114 in 40Mhz bandwidth with 5G mode. As shown in Figure 3.2, the frequency information of signal propagation describes the static activities to some extent. We also find that as long as the layout of the room is fixed, this characteristic is pretty robust.

Therefore, the above results indicate that CSI is a fine-grained reflection of the surrounding environment. The shape of CSI curve can indicate human presence. The subcarrier curve keeps stable under static circumstances but the occupancy presence affects it to some extent.

3.4.2 Methodology

As mentioned in Section 3.2, we can know the situation of human presence by CSI subcarrier curves as seen in the example in Figure 3.2. Through further validation, the curve without any denoising is stable, so we only need to explore a method to compare the shape similarity between adjacent CSI curves in time series. Hence, we employ signal tendency index (STI) [112] to compare the shape similarity of CSI curves. It owns the invariance of translation and scale so that it can compare the tendency and waveform from the statistical point of view. Suppose that $H_1^t, H_2^t, \dots, H_n^t$ is a CSI vector at timestamp t . STI superimposes these curves by translation and scaling operations. The translation step produces

$$H_1^t - \bar{H}^t, H_2^t - \bar{H}^t, \dots, H_n^t - \bar{H}^t, \quad (3.3)$$

where $\bar{H}^t = \frac{1}{n} \sum_{i=1}^n H_i^t$. Then comes the scaling step as follows:

$$\widehat{H}^t = [H_1^t - \bar{H}^t, H_2^t - \bar{H}^t, \dots, H_n^t - \bar{H}^t] / \sigma(H^t), \quad (3.4)$$

where $\sigma(H^t) = \sqrt{\frac{1}{n} \sum_{i=1}^n (H_i^t - \bar{H}^t)^2}$. The transformed object $\widehat{H}^t$ can be used for similarity comparison. As we can derive $\widehat{H}^{t-1}$ in the same manner, it is simple to compute STI to compare two shapes of vectors:

$$S^t = \left\| \widehat{H}^t - \widehat{H}^{t-1} \right\|_2, \quad (3.5)$$

where $\|\cdot\|_2$ denotes the Euclidean norm. Small S suggests that the two curves are similar while large one indicates that they are uncorrelated. As shown in Figure 3.3, the STI values with human presence are much higher than those under the unoccupied situations. We can infer that the shape similarity of CSI curves indeed makes a distinction between occupied and unoccupied situations. An empirical threshold $th_d = 0.8$ selected through cross-validation is introduced to classify the situations. The whole algorithm of occupancy detection is described in Algorithm 1. We implement the prototype on the IoT device and demonstrate that each time of similarity computing can be completed at a frequency of 10Hz, which is sufficient for occupancy detection. The detailed evaluation in Section 3.6 validates its robustness and effectiveness.

Algorithm 1 Occupancy detection using STI

```

1: while IoT device is running well do
2:   Obtain CSI measurements  $H^t$  at timestamp  $t$ 
3:   Calculate the STI value  $S^t = \left\| \widehat{H}^t - \widehat{H}^{t-1} \right\|_2$ 
4:   if  $S^t > th_d$  then
5:     Send result of human presence to MQTT server
6:   else
7:     Continue
8:   end if
9: end while

```

3.5 Case Study: Fine-grained Sensing for Sedentary Behavior Monitoring

To demonstrate the fine-grained sensing enabled by our system, we conduct another case study: leveraging CSI-based sensing for sedentary behavior monitoring that includes static and dynamic human activity recognition.

3.5.1 Sedentary Behavior

Sedentary Behavior (SB) is defined as participation activities that do not substantially increase energy expenditure, mainly referring to sitting. In the modern urban life, SB has become more and more common due to the nature of work, such as clerical work or long periods of driving. Previous research shows that the office workers sat an average of 66% of their workday [113]. Meanwhile, people spend leisure time watching TV and kids spend much time doing homework, which also leads to SB unconsciously [114]. This imperceptible behavior has been proved to be associated with reduced health outcomes and even chronic diseases, including overweight, cholesterol, blood pressure, diabetes and cancer [115]. Apart from the medical research, the data of sedentariness also plays an important role in predicting conditions of patients [116]. Therefore, it is imperative to build a ubiquitous sedentary behavior monitoring system, which can record the duration of SB as well as various activities during it and offer advice when necessary.

Despite the significance of sedentary behavior detection, there does not exist such a ubiquitous and easy-to-use SB monitoring system. Existing methods mainly rely

FIGURE 3.4: Multi-path effect in wireless signal propagation and various daily activities in sedentary behavior.

on kinds of sensors and cameras. The most common one is the accelerometer [113] equipped in wearable devices. Using these devices, medical researchers are able to obtain SB data easily from their experimental subjects, but it is not convenient for ordinary people to use. Thermal sensors [117] are utilized as a SB detection method in workplaces but its user scenarios are limited by the introduction of uncommon sensor. As for computer vision (CV) based method, many state-of-the-art activity recognition algorithms are proposed [118], which may provide high accuracy for SB detection. Nevertheless, the video surveillance may arouse privacy concerns in office and home. It is also highly restricted to line-of-sight (LOS) environment, which brings about loss of efficacy in face of occlusion.

Considering the superiority of WiFi-based methods, we take the first attempt to leverage the COTS WiFi routers to detect sedentary behavior. However, it is quite challenging compared with existing activities recognition systems. Firstly, existing approaches mostly utilize the fluctuations of wireless signals caused by human motions while people may perform relatively static in sedentary behavior. In this case, we cannot directly adopt time sequence information to analyze sedentary behavior. Secondly, most existing methods search a start and end point of a probable signal and match it to the training set via machine learning approaches. For real-time SB monitoring, it may be hard to find such a recognition period. Thirdly, most current methods are designed to detect a specific activity, such as gait recognition [29] and

FIGURE 3.5: Different trends of subcarrier clusters in CSI stream.

people counting [119], while sedentary behavior consists of a random combination of a series of motions such as typing, reading and taking notes shown in Figure 3.4.

To address these problems, we investigate the characteristics of CSI and develop a novel device-free sedentary behavior monitoring system, namely CareFi. The objective of CareFi is to accurately identify whether the user is in SB state by monitoring all possible activities and to record the duration time. At first, we categorize the environment to static and dynamic states. The dynamic activities can be captured by an innovative foreground extraction approach, which models the background using Kernel Density Estimation (KDE). On the contrary, others are defined as static states. Then, dynamic activities are recognized by the features in time domain while static state is more challenging. Inspired by passive indoor localization techniques, we extract the difference among subcarriers to construct posture fingerprints. To deal with the mutable issue of working table, we update the profile of the static states after every dynamic activity.

3.5.2 How Sedentary Activities Affect CSI

Existing CSI-based methods rely on repeatable motions and consequently similar variations of signals. However, in sedentary behavior, there exists a random combination of series of activities. It is not clear that how these factors impact the wireless signals and how CSI reflects their characterizations. Hence, we study the properties of kinds of sedentary activities and conduct preparatory experiments to obtain necessary observations. By analysis of both, we deeply seek for the connections between activities and CSI.

3.5.2.1 Activities in sedentary behavior

Generally, sedentary behavior can be categorized into dynamic and static activities. They have different properties in terms of trajectory, velocity and range of human actions.

- **Static activities:** Static activities, such as sitting still, are very common in sedentariness. People only conduct slight actions during static activities, causing few changes to environment.
- **Dynamic activities:** When people are busy with work or surfing the Internet in leisure time, they usually have some movements to type, write or drink during sedentary time. Typing while sitting should also be considered as sedentary activity, but it is dynamic activities, which should be addressed in our design. Normally, they have different features with regard to velocity and trajectory of actions.

For static activities, *we need to distinguish them from standing still and absence*, since both of them also perform relatively static motions in the environment. On the other hand, our system is supposed to *recognize and record kinds of dynamic activities during a sedentary period*, which can be intended for basic data for medical research. By means of combination of two recognition results, the overall time of SB is used to remind the users to exercise.

3.5.2.2 Connections of Activities and WiFi data

To link the variations of CSI to sedentary behavior, we conduct an experiment to study the characterization. We deploy two TP-Link N750 routers, each with one antenna, as transmitter (TX) and receiver (RX). The transmitter generates and sends the packet every 2 milliseconds on average and the same for RX. The distance between TX and RX is 2 meters and we let the sedentary activities take place in NLOS path with 1 meter away from LOS. Our system can capture 114 subcarriers in 40MHz band at each time stamp. Here comes our observations and discussions:

Different subcarrier information $\rightarrow$ different parts of human body: In each TX-RX pair shown in Figure 3.5, we find that adjacent subcarriers have similar values. However, there often exist multiple different trends of CSI values among subcarriers, represented by three colors. The subcarriers #11-20 with mauve color differ from two other clusters dramatically. The reason behind these observations is that different subcarriers are sensitive to different parts of human body. For example, the movements of arm and leg make values of different subcarriers fluctuate. In this case, the similar values and trends between adjacent subcarrier are redundant information while multiple trends behind CSI stream are valid information. This indicates that we are supposed to extract all clusters of CSI variation trends, which preserves the maximum information of CSI. Inspired by this, we employ principal component analysis (PCA) to obtain significant trends in the design of CareFi.

Time domain $\rightarrow$ dynamic activities: In the analysis of CSI data in time domain, we compare several activities in sedentary behavior such as *sit down*, *stand up* and *type*, shown in Figure 3.6. These activities impact 114 subcarriers in different degrees shown in the visualization of CSI, and we choose a representative subcarrier to display detailed variations. In the time domain, *sit down* and *stand up* cause drastic changes to CSI value while the CSI values of *type* are a combination of many waveforms that may refer to different keystroke knocking according to [34]. To segment the dynamic CSI part, we can derive naive methods based on variance or model the environment using advanced approaches. To distinguish these dynamic activities, we can utilize PCA to obtain impactful trends and then extract critical features such as variance and kurtosis to carry out the classification.

Frequency domain $\rightarrow$ static postures: Although dynamic activities vary obviously in time domain, static activities, such as the beginning part in *stand up* in Figure 3.6, cannot be described by characteristics of fluctuation. The amplitude value difference between them may be caused by specific environment and occlusion, which leads to different propagation paths. Enlightened by multi-path effect and previous indoor localization work, we investigate the subcarrier difference of static activities shown in Figure 3.7 and find the key to the distinction. The static postures and occupancy positions are reflected by the frequency information among subcarriers, which is used as fingerprint for static activity. We also evaluate the robustness of this fingerprint in a month and it does not vary too much in a normal office environment. The next challenging problem is that, as the layout of our table

FIGURE 3.6: The CSI sequences of various dynamic activities.

FIGURE 3.7: Subcarrier difference among static activities.

changes, the fingerprint will also be impacted. However, we know that before each change, there exists a motion which can be captured by CSI fluctuations, and then it is reasonable to update the fingerprints. This is the motivation of profile update introduced in Chapter 3.5.3.

FIGURE 3.8: Framework of CareFi.

3.5.3 Methodology

3.5.3.1 System Overview

The whole framework of CareFi is shown in Figure 3.8, which consists of four main phases including coarse recognition, static posture identification, dynamic activity recognition and profile update. Firstly, our system monitors the CSI stream and uses a coarse recognition method to classify the current state to dynamic or static activity. Then, after noise removal, different categories are processed by different recognition approach. After every time of dynamic recognition, the profiles of static postures need to be updated. During real-time monitoring, CareFi is able to record the duration and activities of sedentary behavior, and recommend the user to exercise if the SB time exceeds a threshold T_e . We will introduce the detailed design of each step in the following section.

3.5.3.2 Coarse Recognition by Motion Extraction

CareFi monitors the CSI stream and obtains a state at each timestamp, including *static state* and *dynamic state* such as sitting and typing respectively. Then, each state can be processed by their corresponding approach. In order to distinguish two states, we need to extract the human motion from raw CSI data with severe interference of noise. As a result, we cannot inherit the variance-based method. Nevertheless, the fact that the background noise on a single subcarrier follows a Gaussian distribution has come to light, and we observe it in our fine-grained CSI shown in Figure 3.9. Using this well-known knowledge, Smokey [30], has succeeded in extracting the motion by Gaussian Mixture Model of background noise. However, too many empirical parameters are required to be tuned, which hinders it for practical implementation. Learning experience from computer vision, we propose a foreground detection method based on kernel density estimation (KDE).

For each subcarrier, we easily have N_b background samples denoted as B that are used to initialize. Let $x(t)_{i,j}$ be the j -th sample of i -th pixel in background B . Using these background samples, we can obtain an estimate of the pixel CSI value pdf at any value using kernel density estimation. Given the observed CSI $x(t)_i$ at time t , we can estimate the probability of this observation as:

$$Pr(x(t)_i) = \frac{1}{N_B} \sum_{j=1}^{N_B} K_\sigma(x(t)_i - x(t)_{i,j}), \quad (3.6)$$

where K_σ is a kernel function with bandwidth σ . Naturally, we choose Gaussian to be our kernel function, and accordingly, the estimated probability can be derived by:

$$Pr(x(t)_i) = \frac{1}{N_B} \sum_{j=1}^{N_B} \frac{1}{\sqrt{2\pi}\sigma_i} e^{-\frac{(x(t)_i - x(t)_{i,j})^2}{2\sigma_i^2}}. \quad (3.7)$$

In this case, the pixel is classified as a foreground pixel if $Pr(x(t)_i) < th_i$, where th_i is a global threshold that controls the percentage of false detection. The main factor of the model is the choice of suitable kernel bandwidth σ_i . A small one may lead to a ragged estimate while too big will neglect some useful changes.

To estimate the kernel bandwidth σ_i , we introduce the method developed in image

FIGURE 3.9: Noise of each subcarrier follows a Gaussian distribution independently.

FIGURE 3.10: Comparison of raw and denoising data on subcarrier #20 in CSI stream.

processing work [120]. After the median m_i of $|x(t)_{i,j} - x(t)_{i,j+1}|$ in the background samples is calculated, assuming this distribution obeys Gaussian distribution $N(\mu, \sigma_i^2)$, the distribution for the deviation $(x(t)_{i,j} - x(t)_{i,j+1})$ is deduced to be a normal distribution $N(0, 2\sigma_i^2)$. Then the median of the absolute deviation m_i is equivalent to the quarter percentile of the deviation distribution, which is

$$Pr(N(0, 2\sigma_i^2) > m_i) = 0.25 \quad (3.8)$$

The standard deviation of the distribution of the i -th pixel can be estimated as

$$\sigma_i = \frac{m_i}{0.68\sqrt{2}} \quad (3.9)$$

which is the kernel bandwidth.

On the whole, we process every frame $\mathbf{h}_t$ by KDE. Once a foreground pixel is detected at time t by $Pr(x(t)_i) < th_i$, we regard this frame as a dynamic frame,

otherwise, the reverse. With a sliding window of N_w and step size of 1 second , we can get the proportion of dynamic and static frames in this period. If the dynamic frames exceed $\rho = \frac{1}{10}$, we affirm that the CSI within this sliding window in dynamic state, otherwise, static state.

3.5.3.3 Data Preprocessing

The coarse classified CSI data matrix $H^t \in \mathbb{R}^{(N_T \times N_R \times 114) \times N_w}$, which is denoted by $\mathbf{H}^t = [\mathbf{h}_{t-N_w+1}, \mathbf{h}_{t-N_w+2}, \dots, \mathbf{h}_{t-1}, \mathbf{h}_t]$, is then preprocessed by noise removal techniques for higher level recognition. We eliminate the inherent noise in both time and frequency domain. The comparison before and after data preprocessing is shown in Figure 3.10. Here come the steps of denoising:

Interpolation: The transmitter suffers from packet loss and lack of priority for high-level task in OpenWrt, so the receiver cannot be guaranteed to work with the same frequency. To deal with this problem, we utilize *linear interpolation* to get a uniform time sequence for CSI data.

Subcarrier refinement: Due to multi-path effect, some signals propagate with severe distortions after reflection at a far distance, which leads to the fluctuations of CSI in a rambling manner. These signal paths cannot reflect human motion and have a longer delay profile than useful ones. Thus, for each frame of CSI data, we use Inverse Fast Fourier Transform (IFFT) to get channel impulse response (CIR) and remove the components whose delay goes beyond 0.5 ms. Then we convert CIR back to CFR by Fast Fourier Transform (FFT).

Temporal de-noising: In modern wireless communications, WiFi devices can adjust transmission power to ensure highest transmission capacity, which brings about high frequency noise in the time domain of CSI. We choose a wavelet-based de-noising approach to remove it. Wavelet transform uses various transforms $\psi_{a,b}(t) = a\psi(t-b)$ obtained from a basic function through scaling and translation in terms of parameter a and b respectively. The wavelets coefficients of our CSI time sequence $r(t)$ can be calculated by:

$$W(a, b) = \int_{-\infty}^{\infty} \psi_{a,b}^*(t) r(t) dt \quad (3.10)$$

where $*$ is the complex conjugate and *Symlet* is chosen to be the basic function ψ . After the Discrete Wavelet Transform (DWT) on the CSI data, we filter the wavelet coefficients and get the sanitized CSI data by the inverse transformation.

3.5.3.4 Dynamic Activity Recognition

As illustrated in Section 3.5.2, we should analyze the principal component of CSI and derive discriminative features from it. After normalization, PCA is used to decompose the dynamic CSI data $\mathbf{H}_t$ and we find that first four components take up 99.8% information of wireless signals. In Figure 3.11, the CSI records four human motions of sitting up or down, reflected by four peaks of PCA. Furthermore, we can see that different PCA components are sensitive to different variations in CSI. The third component is sensitive to the first two actions but the fourth component reflects the second and fourth motion. This allows us to extract effective features from these components, which reduces the dimension of feature vectors and complexity of computation.

In signal processing, there exist hundreds of methods to depict a signal property. To choose suitable features in the time domain, we calculate the information gain during classification, and select eight feasible features: *entropy*, *mean*, *variance*, *standard deviation*, *kurtosis*, *skewness*, *inter quartile range* and *mean crossing rate*. Then, for each H_t^d , we have a dynamic feature vector $\mathbf{v}_d \in \mathbb{R}^{N_R \times N_T \times 4 \times 8}$ for classifier. We compare classic state-of-the-art classifiers to choose the best one referring to *discriminant analysis classifier*. This helps CareFi to efficiently distinguish kinds of common dynamic activities such as *typing* and *writing*.

3.5.3.5 Static Posture Identification

Former observation elaborates the correlation between frequency diversity and static posture, so subcarrier difference is considered as fingerprint of static posture. When we need to classify a period of CSI data H_t , we are supposed to compare it with all possible fingerprints of postures, which requires to be calibrated in a new environment for the first time. Hence, each posture owns a fingerprint obtained by averaging β frames. As each undefined CSI data has N_w frames, we firstly obtain

FIGURE 3.11: Top 4 principal components of CSI using PCA.

the posture of each frame by comparison, and count the proportion of each posture. The largest one indicates the most probable posture.

To efficiently compare the fingerprint of each frame $h_i \in \mathbb{R}^{114}$, $i \in [1, N_w]$ and a posture template h_p , we introduce Pearson product-moment correlation coefficient (PPMCC) which varies from -1 to 1 . The absolute of PPMCC is a measurement of similarity between two signal vectors. The larger result indicates more similarity. In mathematics, PPMCC owns the invariance of translation and scale so that it can compare the tendency and waveform from the statistical point of view. The PPMCC ρ between h^i and h^p is defined as the quotient of covariance divided by their standard deviation:

$$\rho_{h^i, h^p} = \frac{\text{cov}(h^i, h^p)}{\sigma_{h^i} \sigma_{h^p}} = \frac{E[(h^i - \mu_{h^i})(h^p - \mu_{h^p})]}{\sigma_{h^i} \sigma_{h^p}}. \quad (3.11)$$

When we estimate the covariance and standard deviation of the sample space, we obtain the sample correlation coefficient r calculated as:

$$r_{h^i, h^p} = \frac{\sum_{j=1}^{N_w} (h_j^i - \bar{h}^i)(h_j^p - \bar{h}^p)}{\sqrt{\sum_{j=1}^{N_w} (h_j^i - \bar{h}^i)^2} \sqrt{\sum_{j=1}^{N_w} (h_j^p - \bar{h}^p)^2}}, \quad (3.12)$$

where $\bar{h}^p$ and $\bar{h}^i$ are the mean of two vectors respectively.

FIGURE 3.12: The layouts of the dormitory and office room.

3.5.3.6 Profile Update and Close Loop

In the experiment of CareFi, we find that the fingerprint of posture can easily get a bit difference by the change of table layout, such as putting a book or bag on the working table. This impact for once will not lead to false recognition but accumulation effect can change the fingerprint totally. Therefore, we update the fingerprint of *sitting*, *standing* and *empty* after each dynamic activity happens. The new detected profile of posture is used to replace the initial one as an effective update. Although this may cause a mistake sometimes, it works well in most cases as the accuracy of posture identification is very high. When false situation happens, the system needs manual labeling of the right posture. Besides, as our system has sufficient sampling rate of CSI measurements, a new profile can be easily updated using the average of new profile $\hat{H}_t$. So far, we have constructed a close loop in our system. CareFi monitors the CSI stream with a sliding window. Coarse recognition classifies the current CSI to either dynamic activity or static posture. For dynamic activity, CareFi will recognize the activities during sedentariness for medical research data, and update the profile of postures. For static posture, CareFi will distinguish three states. Both categories of activities caused by human will be recorded for the duration of sedentary behavior. Finally, if the SB time exceeds a threshold, CareFi will remind the user to exercise.

3.6 Evaluation of Occupancy Detection

We implement a prototype of our IoT platform based on two TP-LINK N750 wireless dual band routers as IoT devices, and a cloud server as computing center.

Two routers, used as TX and RX, are upgraded to our innovative OpenWrt system so that CSI is reported on board directly. Note that previous work [121, 122] also employ this mode of access point. Nevertheless, they leveraged normal OpenWrt firmware without any modifications in the physical layer. In our architecture, TX with one antenna was set to 802.11n AP mode at 5GHz with a 40MHz bandwidth and the client, RX, with 3 antennae, was connected to this WiFi network. All the routers are deployed at a height of 1.5m. In this case, multiple clients are allowed in our system, which can form several pairs of network and increase the coverage. Hence, the number of CSI streams is $1 \times 3 \times 114$ and we prototype our occupancy detection and human activity recognition system based on this platform. We conducted the evaluation in an office zone ($5.6m \times 9m$) and a conference room ($5m \times 7m$).

For a coarse sensing task with the low complexity of Algorithm 1, we implement the occupancy detection function directly on IoT devices. It can detect the presence of any number of occupants. The source codes are written in C languages and they are cross-compiled for the OpenWrt system. We evaluate CSI-enabled occupancy detection in terms of accuracy, computing speed and antenna usage. For the evaluation of accuracy, we use three metrics:

- Overall Detection Accuracy (ODA):

$$ODA = \frac{\text{corrected detected samples}}{\text{all samples}}$$

- Occupied Detection Ratio (ODR)

$$ODR = \frac{\text{detected occupied samples}}{\text{true occupied samples}}$$

- Unoccupied Detection Ratio (UDR)

$$UDR = \frac{\text{detected unoccupied samples}}{\text{true unoccupied samples}}$$

As shown in Table 3.1, we evaluate the accuracy using 3 single antenna as well as all antennae. It is obvious that when we use CSI of only one antenna, the computing time is low but the accuracy only reaches 91%-94%. Then with CSI data of three antennae (3×114 subcarriers), the accuracy runs up to 96.8% and the computing

TABLE 3.1: Occupancy Detection Performance

Antenna	Frequency (Hz)	ODA (%)	ODR (%)	UDR(%)
RX1	30	93.2	95.2	92.4
RX2	30	91.5	93.1	89.6
RX3	30	94.8	96.3	93.8
RX 1,2,3	22	96.8	98.2	95.1

frequency is about 22Hz which is fairly acceptable. We also discover that the ODR is a little higher than UDR under all circumstances. It is because our method is more sensitive to human motion in line-of-sight (LOS) and has a limited range. When human motions occur around the range border, it is hard to distinguish the presence. This results in some wrong detection of unoccupied situation. In summary, based on all subcarriers of three antennae and our IoT devices, our CSI-enabled occupancy detection achieves high accuracy with satisfactory frequency. The notification of presence is then forwarded to the user end by our cloud server.

3.7 Evaluation of Human Activity Recognition

We implement a prototype of CareFi with TP-LINK N750 wireless dual band routers. Having upgraded their firmwares, we can directly obtain the CSI measurement from the routers. The transmitter with one antenna ($N_T = 1$) was set to IEEE 802.11n AP mode on Channel 36 at 5GHz with 40MHz bandwidth, while the receiver with three antennae ($N_R = 3$) was connected to this WLAN network. We configured the receiver to record the CSI with a sampling rate of 500Hz. Then the receiver sent raw data to our server by UDP protocol and the methodologies were implemented on the server. Thus, the system was a real-time monitoring system with algorithm implemented in the server and CSI transmission from access points. The programs running on the routers were written by C languages and the algorithms were implemented by MATLAB. We evaluated the performance of CareFi in two common environments: (a) a dormitory of $4.0 \times 3.8m^2$ and (b) an office room of $5.0 \times 6.0m^2$ in an office building. Their layouts and the positions of routers are shown in Figure 3.12. Five volunteers were engaged in the experiment to test the robustness of recognition, and they were asked to sit in the chair between RX and TX in the both layouts. The evaluation in the office room was conducted in work time when other staff were working around in their cubicles, while the experiment

FIGURE 3.13: ROC curves of CareFi in LOS/NLOS situations.

in the dormitory room tested the single-user case. The system was assessed from four perspectives.

3.7.1 Overall Evaluation

To simulate the real-world cases, we use NLOS configurations of routers, and run CareFi during workdays to investigate the duration of sedentary behavior of users. The ground truth is provided by the record of volunteers and the threshold $T_e = 30$ indicates that CareFi will alarm when the duration of SB reaches $30mins$ (allowable error $\pm 1min$). To scientifically assess the performance, we consider two reasonable evaluation criteria: (1) True Positive Rate (TPR): the proportion of cases where CareFi correctly reminds the user on time and (2) False Positive Rate (FPR): the fraction of cases where CareFi mistakenly reminds the users when the record is not right. On the whole, we conduct one hundred sets of experiments of different user in 3 workdays and CareFi successfully sent correct alarms by 97% with the TPR of 0.979. To quantitatively compare the accuracy in two environments, we plot the Receiver Operating Characteristic (ROC) curves of CareFi in Figure 3.13 that indicate the tradeoff between TPRs and FPRs. We can see that NLOS case in dormitory represented by red line achieves better performance than the case in office represented by yellow line. Given the FPR of 0.01, the TPRs are 0.941 and 0.873 for the two cases. This may be caused by more environmental noise in office.

Apart from the duration monitoring, we also evaluate the dynamic activity recording during this period. We evaluate six common activities including *read*, *write*, *type*, *drink*, *sit up or down* and *phone*, as well as environmental interrupt which is

FIGURE 3.14: Confusion matrix of dynamic activity recognition.

also recorded by coarse recognition. In Figure 3.14, the confusion matrix displays an average accuracy of 94.9%. The accuracy of classification between environmental interference and human activities achieves 97.5% and others have a reasonable accuracy between 90.5% to 100.0%.

3.7.2 Impact of LOS/NLOS Circumstances

To deeply understand the impact of LOS and NLOS cases, we conduct the comparative test in the dormitory. In Figure 3.13, the case of LOS achieves better accuracy than the NLOS case. However, when the given FPR is lower than 0.025, NLOS performs better. The reason is that the LOS case is too sensitive to motions happening in the LOS, so even slight actions can interrupt the identification of postures. We investigate the posture identification in two cases shown in Figure 3.15, which proves our guess.

To better demonstrate the effectiveness of our posture identification method with profile update, we compare it with two state-of-the-art approaches, E-eyes [33] and CARM [123] in both LOS and NLOS cases. As shown in Figure 3.16, our approach performs better for sitting and standing in LOS situation but gets a similar result for empty room. This is also because the LOS is too sensitive to motions in our approach, but other two approaches use either a threshold or a transition motion to

FIGURE 3.15: Confusion matrix of posture identification in LOS and NLOS situations.

distinguish whether a person exists. In contrast, their designs show flaw in NLOS case. Since the recognition of motion such as sitting up is less accurate due to the signal characteristics, the wrong classification of transition motion may lead to wrong posture identification. Our approach obtains the best result in NLOS situations for all postures.

3.7.3 Impact of Environmental Dynamics and People Around

For the robustness of CareFi, we design an experiment to test the performance of posture identification during 20 days. It should validate that the profile update keeps the fingerprints of static postures adaptable and we set the update time length to be 2 seconds. In this test, we changed the layout of the table between RX and TX every day. For instance, we put some textbooks on the table on the 2nd day and moved the computer monitors on the 9th day. As shown in Figure 3.17, the accuracy with profile update keeps steady but the one without updating scheme drops dramatically from 97.2% to 19.7% after the 10th day. The profile update addressed the problem by learning new fingerprints after changes, and thus it achieved better results. This proves the feasibility of profile update and the robustness of CareFi against dynamic environment.

We also find that the larger the distance between TX and RX, the larger the coverage of CSI sensing. In the dormitory, there mostly exists one person. Then we put the TX and RX far from each other to enlarge the sensing coverage. In the office, we put TX and RX near each other, and thus the people around cannot make considerable effects on CSI. In the real-world experiment, people around in

FIGURE 3.16: Comparison of posture identification accuracy in LOS and NLOS situations.

FIGURE 3.17: Overall performance of posture identification with or without profile update.

the office do not influence our system at all because they are out of the sensitive coverage of CSI and only the user activity dominates the signals.

3.7.4 Comparison of Rough and Fine-grained CSI

As mentioned before, our system is purely based on routers with fine-grained CSI data which firstly supports 114 subcarriers in 5GHz rather than 30 subcarriers in other existing CSI sensing systems. To prove more information hidden behind, we simulate the 30 subcarriers case according to Intel 5300 NIC tool [24], because the current systems mostly employ existing devices with Intel 5300 NIC as hardware infrastructure. In Figure 3.13, the case of 30 subcarriers achieves the overall accuracy of 91.2% but given the FPR of 0.01, the TPR drops to 0.647, which is much lower

than 0.941, the case of 114 subcarriers. The reason is that some activities cannot be captured with less subcarrier number. This comparison shows that our system with more subcarrier information enhances the robustness and accuracy. Furthermore, the results also demonstrate that our firmware contributes to improving other existing WiFi-based activity recognition systems from the perspective of fine-grained CSI data.

3.8 Conclusion

In this chapter, we design a CSI-enabled IoT platform for occupancy sensing in smart homes, and present two innovative applications based on it. Our platform only employs commodity WiFi routers and thus can be easily deployed in smart homes. It consists of three parts: IoT platform, cloud server and user end. We firstly integrate OpenWrt system with CSI tools and then the CSI reported in the IoT devices will be transmitted to cloud server for processing. Finally the results are displayed in the user end. To demonstrate the performance of our system, we implement two cases including occupancy detection and human activity recognition. We present the coarse case, occupancy detection, by means of exploiting the CSI curve features in terms of the presence of persons. It is completely realized in the IoT device and the evaluation shows an accuracy of 96.8% with 22Hz. For more fine-grained case, we propose CareFi, a device-free sedentary behavior detection system only using COTS WiFi routers. We investigate the characteristics of CSI under various circumstances and find the discriminative features of both dynamic and static activities. To distinguish these two categories, we develop a motion detection method based on kernel density estimation. Then dynamic activities can be recognized by features in time domain while static postures are identified by frequency diversity in CSI. To improve the robustness to the environment, we leverage profile update after each possible human motion. In the experiment, we prototype CareFi on our system. Extensive real-world experiments have been conducted and they validate that CareFi achieves an outstanding sedentary behavior recognition performance in terms of accuracy, robustness and feasibility.

Chapter 4

Robust Representation Learning for Few-shot WiFi-based Gesture Recognition

In this chapter, following the work in Chapter 3, a harder case, WiFi-based gesture recognition, is investigated. Section 4.1 introduces the current gesture recognition methods and the motivation of this chapter. Section 4.2 briefly formulates the CSI data and how the data is associated to gestures. In Section 4.3, a Siamese architecture is proposed to extract useful patterns from noisy CSI data. Extensive experiments are conducted in Section 4.4 and the whole chapter is concluded in Section 4.5.

4.1 Introduction

Gesture recognition plays a vital role in the thriving research field of human-computer interaction (HCI). It is an underpinning technology for a wide range of applications including but not limited to wearable devices, family entertainment systems and smart driving. For example, Xbox applies gesture recognition to emulate sports games, while BMW has developed gesture-based remote control in its automotive infotainment system. Furthermore, precise gesture identification can be considered as the first step towards automatic sign language recognition, which

would greatly facilitate communications with people having hearing and speech difficulties.

Compared with other IoT sensors, WiFi modules are more common in existing smart appliances, and thus the communications between appliances and the WiFi routers form multiple pairs in a smart home, enabling WiFi-based indoor sensing technology. Recent advances in CSI have demonstrated its ability to capture adequate sensing information for indoor localization [26], activity recognition [124–126] and even human identification [29] in a device-free, privacy-preserving, and cost-effective manner. Considering its pervasive availability and high accuracy, CSI is a promising choice for gesture recognition.

The key issue with any CSI-based application, however, is that the measurement is never clean in real-world settings. When it comes to CSI-based gesture recognition, the measured signal could be heavily distorted by walls, object placement, individual heterogeneity, etc. Those interferences, although might be useful for other purposes, are irrelevant to gesture recognition, leading to two challenges that we strive to resolve in this work. (1) From a statistical learning perspective, a robust representation must be learned to simultaneously capture the intrinsic, gesture-related spatio-temporal feature and remove underlying environmental noises. The imputed noise in CSI data is structured and may also contain discriminative information when limited samples are collected. Hence, a direct use of off-the-shelf deep learning techniques or traditional classifiers with hand-crafted features will be significantly impaired by the issue of overfitting. (2) From the application perspective, a gesture recognition system is supposed to be transferable across peripheral conditions. In other words, the system should also work well in a new environment with minimal tuning. Specifically, very few training samples should be required or collected to adapt the system to a new environment. This is hard to achieve via conventional deep neural networks.

In this chapter, we propose a WiFi-based gesture recognition system that leverages the CSI measurements from WiFi devices and identifies the gestures via a novel Siamese recurrent convolutional architecture. The framework of our recognition system is illustrated in Figure 4.1. Addressing the challenges mentioned above, our contributions are highlighted as follows:

- For gesture recognition, we learn a robust gesture-related representation by developing a novel deep model, which adopts a convolutional neural network (CNN) to extract spatial features, and coupled with a recurrent neural network (RNN) to capture temporal dynamics of gestures.
- A Siamese architecture with transferable pairwise loss is proposed to clean up other structured but irrelevant features such as environmental dynamics. The net effect is that only the features relevant to our recognition task are retained. More interestingly, the proposed learning framework also enables one-shot learning for high availability in practical use, as detailed later.
- The experiments are conducted in multiple real-world environments, and the results show significant improvements over existing approaches.

FIGURE 4.1: Illustration of our Siamese recurrent convolutional architecture. The model accepts a new gesture input and extracts its deep features, which are then compared with all the historical one-shot templates to choose the most similar one.

4.2 System Design

4.2.1 IoT Platform for Sensing Gestures

In indoor environments, WiFi signals propagate from TX, such as access point, to RX, such as mobile devices through multiple paths in virtue of reflection, scattering and diffraction caused by walls, furniture and movements of human bodies [27]. At present, most of the commercial off-the-shelf (COTS) WiFi routers are tailored for multiple input, multiple output (MIMO) communication with several antennae. Orthogonal frequency-division multiplexing (OFDM) is commonly adopted, supporting IEEE 802.11n/ac standard. The data rate is increased by sending packages via a mass of narrow subcarriers at different frequencies simultaneously. Therefore, CSI contains amplitude attenuation and phase shift of multiple paths in each subcarrier. The details of the WiFi-based sensing system and the CSI data have been introduced in Chapter 3. The platform reports all the 114 subcarriers for the 40MHz bandwidth on each TX-RX pair operating on 5GHz. Let N_{TX} and N_{RX} denote the number of TX and RX antennae respectively. At each timestamp, our platform reports $N_{TX} \times N_{RX} \times 114$ CSI streams, and naturally, higher dimensional data provides more fine-grained sensing information.

4.2.2 Insights: Learning Gestures from CSI Phase

Existing CSI-based approaches for gesture recognition are dependent on the amplitude of CSI streams [127], because phase information appears disorderly due to channel distortion. In our preparatory experiment of CSI data, we discover that *phase difference* is more sensitive to environment changes. Since the variance of phase difference across two antennae stands for the sum of individual variance on each antenna, the CSI phase difference is a better choice for gesture recognition. As shown in Figure 4.2, we can find a similar pattern of the same gesture across time in one CSI stream, and all subcarriers at a time instance characterize the multi-path effects in the current environment. The observations above imply that the CSI fingerprints of gestures have both spatial and temporal characteristics that are useful for recognition.

FIGURE 4.2: Phase difference of CSI readings, including two kinds of gestures, in which the first two are *downward* and the last two are *leftward*.

4.3 A Siamese Recurrent Convolutional Network Architecture

In this section, we provide a detailed introduction to the proposed Siamese architecture. We firstly introduce the model for representation learning and the Siamese architecture for one-shot learning. Then, a pairwise loss with transferable metric is proposed and analyzed.

4.3.1 Model

A Siamese neural network consists of two twin networks, whose parameters are tied [128]. They accept distinct inputs but their outputs are joined at the top. Sharing weights of twin networks guarantees that similar inputs can obtain akin outputs in the feature space. Moreover, due to the symmetry of network structure, the order of input pair does not affect the ultimate output in the conjoining layer. As illustrated in Figure 4.1, our model extends the standard Siamese structure by incorporating representation and metric learning for tensor-like, high resolution CSI data. We start with the feature learning in our twin network.

The CSI data for each sample of gesture is denoted as follows:

$$X \in \mathbb{R}^{N \times T}, \quad (4.1)$$

where N is the number of subcarriers and T is the length of the time sequence. Intuitively, the CSI of one gesture is akin to a video clip, where N looks like the pixel in one frame while T is the number of frames. Our learning model firstly learns the spatial feature from each frame by means of CNN. Let $C(n_c, s_k)$ denote

a 1D-convolutional layer with the kernel size s_k and the number of feature maps n_c . The CNN architecture is represented by the shorthand notation: $C(16, 11) \rightarrow P \rightarrow C(32, 3) \rightarrow P \rightarrow C(64, 3) \rightarrow P$, where P is a max-pooling layer. We deploy *leaky rectified linear units* (leaky ReLUs) [129] in each layer, whose activation function is defined as $h : x \mapsto \max(\alpha x, x)$, where α is a parameter learned during back propagation. Since the gradient is zero when the conventional ReLU unit is not active, the weights of some inactive units may never be activated. This is known as the dead ReLU problem [130]. Leaky ReLU reserves part of negative component, and works better in our architecture.

Given the feature map from CNN, we consider the temporal features learning by RNN. The core principle of RNNs is to learn the temporal dynamics by creating internal memory to store them. Traditional recurrent neural networks build up their states from previous time steps, which is quite limited. In our situation, gesture recognition should be sensitively context-aware, because a gesture becomes recognizable after a few frames. Therefore, we employ a bidirectional Long Short-Term Memory (LSTM) model to process a sequence in two directions. The bidirectional LSTM (BiLSTM) computes two hidden sequences including the forward one $h^{(f)}$ and backward one $h^{(b)}$. One standard LSTM sequentially updates a hidden sequence by a memory cell composed of four parts: a memory state c_t , an output gate o_t that controls the effect of output, an input gate i_t and a forget gate f_t that decides what gets stored in the memory. Below are the updates in order to obtain one hidden sequence such as $h^{(f)}$. The LSTM is parameterized by weight matrices $W_i, W_f, W_c, W_o, U_i, U_f, U_c, U_o$ and biases b_i, b_f, b_c, b_o , and the whole update is performed at each $t \in \{1, \dots, T\}$:

$$i_t = \sigma(W_i x_t + U_i h_{t-1} + b_i), \quad (4.2)$$

$$f_t = \sigma(W_f x_t + U_f h_{t-1} + b_f), \quad (4.3)$$

$$\tilde{c}_t = \tanh(W_c x_t + U_c h_{t-1} + b_c), \quad (4.4)$$

$$c_t = i_t \odot \tilde{c}_t + f_t \odot c_{t-1}, \quad (4.5)$$

$$o_t = \sigma(W_o x_t + U_o h_{t-1} + b_o), \quad (4.6)$$

$$h_t = o_t \odot \tanh(c_t), \quad (4.7)$$

where σ is a Sigmoid function. Eventually, we adopt a fully connected layer to obtain a feature map v from the last state of the bidirectional LSTM. We empirically

set the number of the hidden units in BiLSTM to be 64. In general, it is believed that the feature map subsumes both the spatial and temporal features by CNN and BiLSTM respectively.

4.3.2 Transferable Pairwise Loss

Through the Siamese architecture, the feature maps of two gesture inputs are combined at the top. We can use either the traditional contrastive energy function or the cross-entropy objective in the conjoining layer, so as to maximize the distance between samples of different class and vice versa. Suppose that x_i and x_j are two input training samples and we attain two feature vectors $\mathbf{f}(x_i)$ and $\mathbf{f}(x_j)$ respectively via twin networks. Let $\mathbf{y}(\mathbf{x}_i, \mathbf{y}_j)$ be an integer which indicates the label for a pair of inputs:

$$\mathbf{y}(x_i, x_j) = \begin{cases} 1 & \text{if } x_i \text{ and } x_j \text{ belong to the same class} \\ 0 & \text{if } x_i \text{ and } x_j \text{ belong to different class} \end{cases} \quad (4.8)$$

Then we adopt a novel pairwise loss with less constraint than the contrastive loss [131]. The loss function between two samples can be given as follows:

$$\begin{aligned} \mathcal{L}_c(x_i, x_j, \theta) = \\ \max(0, b - (2\mathbf{y}(x_i, x_j) - 1)(m - \mathbf{D}(\mathbf{f}(x_i), \mathbf{f}(x_j), \theta))), \end{aligned} \quad (4.9)$$

where two parameters b, m satisfy $0 < b < m$, θ is the parameters of our network and $\mathbf{D}(x, y, \theta)$ denotes a metric distance between two samples under parameters θ . The pairwise loss guarantees that the distance between the same samples is restricted within $m - b$ while the margin of samples from different classes exceeds $m + b$. To better understand it, the pairwise loss can be rewritten as follows:

$$\begin{aligned} \mathcal{L}_p(x_i, x_j, \theta) = \mathbf{y}(x_i, x_j)\max(0, b - m + \mathbf{D}(\mathbf{f}(x_i), \mathbf{f}(x_j), \theta)) \\ + (1 - \mathbf{y}(x_i, x_j))\max(0, b + m - \mathbf{D}(\mathbf{f}(x_i), \mathbf{f}(x_j), \theta)), \end{aligned} \quad (4.10)$$

where the weaker constraint $b - m$ is more sensible for a real-world problem. The above reformulation is intuitive and matches well in our learning task, i.e., our model should maximize the relevant features and minimize the structural environmental noise.

FIGURE 4.3: How pairwise loss and domain confusion loss work for biased data. Three colors represent different domains and the shapes denote their labels.

To be more specific, imagine that we feed the Siamese network with two gesture samples in different environments, i.e. two domains. During training, the metric learning would minimize the distance between the samples in the same class from the two domains, which is equivalent to maximizing similar characteristics across the two domains. In this way, only common features that are invariant and relevant to gesture recognition will be extracted. If we input a pair of different gesture samples from both the same and different domains, the model will maximize both distances [128], which discards irrelevant patterns that are “noise” to our application. Normally, L_1 or L_2 norm based distance works well for the metric learning, which learns robust representations and semantic features.

Yet another significant problem lies in the transferability. How can we train our learning model to have stronger capacity of domain adaptation? It is well known that if we can draw their distributions closer, then the model can generalize better. Inspired by the recent advances, we find that multiple kernel variants of *maximum mean discrepancies* (MK-MMD) bear the unique property for measuring two distributions of samples from different domains [57]. It is reasonable to adopt MK-MMD in our loss function to confuse different domains for the model, and thus we design our cross-domain loss to get better domain-adapted ability.

It is known that the reproducing kernel Hilbert space (RKHS) is a Hilbert space of functions where point evaluation is a continuous linear functional. Let $\mathcal{H}_k$ denote RKHS endowed with a characteristic kernel k . Suppose that we are given a pair of gesture samples x_i^s, x_j^t from domains $\mathcal{D}_s, \mathcal{D}_t$ respectively and the two domains can be characterized by probability distributions p and q . We can compare the two distributions in RKHS $\mathcal{H}_k$, in which the *mean embedding* of distribution p is the

unique element $\mu_k(p)$ such that $\mathbf{E}_{x \sim p}[\phi(x)] = \langle \phi(x), \mu_k(p) \rangle_{\mathcal{H}_k}$ for all $\phi \in \mathcal{H}_k$. The MK-MMD between p and q is defined as the RKHS distance between their mean embeddings. Its squared formulation is defined as:

$$\mathbf{MMD}^2(p, q) \triangleq \|\mathbf{E}_p[\phi(X_S)] - \mathbf{E}_q[\phi(X_T)]\|_{\mathcal{H}_k}^2, \quad (4.11)$$

where X_T and X_S are two distinct domains, and $\phi(x)$ operates on data points x_s and x_t . The empirical approximation to this distance is obtained as follows:

$$\mathbf{MMD}(X_S, X_T) = \left\| \frac{1}{|X_S|} \sum_{x_s \in X_S} \phi(x_s) - \frac{1}{|X_T|} \sum_{x_t \in X_T} \phi(x_t) \right\|_{\mathcal{H}_k}. \quad (4.12)$$

The characteristics kernel associated with the feature map ϕ , $k(x_i^s, x_j^t) = \langle \phi(x_i^s), \phi(x_j^t) \rangle$, is defined as the convex combinations of m Positive Semi-Definite (PSD) kernels k_u ,

$$\mathcal{K} \triangleq \left\{ k = \sum_{u=1}^m \beta_u k_u : \sum_{u=1}^m \beta_u = 1, \beta_u \geq 0, \forall u \right\}, \quad (4.13)$$

where the constraints on coefficients β_u assure that the multi-kernel k is characteristic. As proved in [57], the choice of kernels for MMD plays a vital role in the test power and error. MK-MMD leverages multi-kernel k_u to enhance the ability of a metric. Empirically, we consider a family of m Gaussian kernels k_u^m with distinct bandwidth γ_u in MK-MMD.

As a result, we introduce a confusion loss via MK-MMD to the loss function under our network parameter θ :

$$\mathcal{L} = \mathcal{L}_{\mathbf{p}}(x_i^s, x_j^t, \theta) + \lambda \mathbf{MMD}^2(X_i, X_j), \quad (4.14)$$

where $\mathcal{L}(x_i^s, x_j^t, \theta)$ denotes the pairwise loss, $\mathbf{MMD}^2(X_i, X_j)$ is the distance between two domains with data X_i and X_j , and λ determines how strongly the model confuses the domains. The loss function is calculated using the result of the final fully connected layer.

The final loss function can be formulated as a multi-task problem illustrated in Figure 4.3. The pairwise loss minimizes the Euclidean distance between the same class and maximizes the distance between different classes. It helps learn robust features independent of background noise. Meanwhile, the domain confusion loss

closes the gap between domains. Our loss enhances the metric learning in the perspective of transferability, which is conducive to learning strong representations, and simultaneously meeting the criteria of domain transfer. After the model is trained with enough samples, we can just collect and annotate one or few samples of each class in a new environment. Then we compare the Euclidean distance $\mathcal{D}(x_{new}^s, x_c^t, \theta_t)$ between the test sample and each class, so that the minimum one indicates the predicted result. L_2 distance is chosen due to empirical results. Hence, both representation learning and one/few-shot learning are achieved simultaneously in this manner. The whole mechanism is summarized in Algorithm 2 for better understanding.

Algorithm 2 One-shot gesture recognition

```

train the network  $f$  and obtain the network weights  $\theta$ 
obtain the features of  $k$  gesture templates  $T = \{f(g_i, \theta) | i < k\}$ , where  $k$  is the
number of gesture category
obtain the feature of the testing sample  $f(g', \theta)$ 
set  $j = 0$ 
repeat
    calculate the  $L_2$  distance  $D(f(g_j), f(g'), \theta)$ 
until  $j < k$ 
output  $y(g') = \operatorname{argmin}_{j < k} D(f(g_j), f(g'), \theta)$ 
  
```

4.4 Experiment

4.4.1 System Settings and Data Collection

We implemented our gesture recognition system using two TPLINK N750 routers (as TX and RX) by upgrading their firmwares to our CSI enabled platform. Each router has three omnidirectional antennae and operates on the 5GHz frequency band in order to reduce the signal interference. CSI data is collected in the 40MHz bandwidth at 5GHz so it has 114 subcarriers for each TX-RX pair. Shown in Figure 4.5, two routers were set to be 0.75m apart and the gestures were performed in LOS. The distance, 0.75m, is selected empirically by visualizing the CSI of the gestures. Under the 0.75m distance, we find that the patterns of CSI are clearer, and shorter or longer distance will reduce the CSI sensing granularity. The routers were put on a wooden table. We collected CSI data with a sampling rate of

40 packages/s, and a linear interpolation was adopted to ensure the stationary interval of consecutive CSI frames. To validate the recognition performance, we introduce six gestures in Figure 4.4: *upward(UW)*, *downward(DW)*, *leftward(LW)*, *rightward(RW)*, *push(PH)*, *pull(PL)*, and assume the duration of each gesture is less than 4 seconds. In this case, the CSI data for each gesture sample is $3 \text{ antennae} \times 114 \text{ subcarriers} \times 160 \text{ frames}$, and thus the size of CSI phase difference is $(2 \times 114 \times 160) = (228 \times 160)$ that is the input size of our network. We recruit 10 volunteers to collect 40 samples for each gesture (totally $10 \times 6 \times 40$ samples) under two different circumstances including a large meeting room ($5\text{m} \times 7\text{m}$) and a small one ($4.4\text{m} \times 6.1\text{m}$), which are shown in Figure 4.5.

FIGURE 4.4: Six gestures used in the experiment.

FIGURE 4.5: The environmental settings of the meeting rooms.

4.4.2 Learning Configurations

Model learning configurations are detailed in this subsection. Note that no data augmentation is utilized because affine distortions are meaningless for CSI data with no invariance to rotation and scale.

TABLE 4.1: Overall accuracy of different methods

Apporach	Accuracy	One-shot Accuracy
WiG	75.8	18.9
WiGeR	81.5	21.0
AAAI-18	83.3	36.6
AlexNet	85.0	-
ResNet-18/50/152	88.6 / 89.5 / 90.0	-
CNN-LSTM (Ours)	92.8	-
SN-MMD (Ours)	-	89.5

Weight initialization. We initialize all network weights in CNN from a normal distribution with zero-mean and a standard deviation of 10^{-2} , while the weights in LSTM follow a uniform distribution in $[-0.125, 0.125]$.

FIGURE 4.6: Confusion matrix of gesture recognition using different methods.

Learning schedule. We follow a standard backpropagation algorithm where the gradient is additive across the twin networks due to the tied weights. Stochastic gradient descent (SGD) is adopted with a batch size of 30, learning rate $\eta_j = 0.001$ and momentum $\mu_j = 0.9$. We sample randomly *same* and *different* pairs for training. The training process starts with a learning rate η_0 and it keeps decaying uniformly across the network by 1 percent per epoch. The training ends when the validation accuracy does not increase by a threshold. In addition, we set $m = 1.5, b = 1.0$ in the pairwise loss.

Hyperparameter optimization. We use a Bayesian optimization framework [132] for hyperparameter selection. We set the optimizer to maximize the one-shot validation accuracy by treating the following hyper-parameters as the variables for Bayesian optimization: initial learning rate, the kernel size in the convolutional layer, the number of recurrent layers, and the number of units in a fully connected layer. The best parameters for our network are shown in Chapter 4.3.1.

4.4.3 Evaluation

Our work contributes to learning robust spatial-temporal features and enables one-shot learning via the Siamese architecture.

Overall performance: We firstly evaluate our model in a traditional way where the training and testing data come from the same person in the same environment. We keep the CNN and LSTM in our network and add a fully connected layer with 6 units as well as a Softmax layer to it. In this way, our representation learning model is extended to a classification model, which can be compared with other state-of-the-art methods. We compare our approach with two advanced methods for gesture recognition. Specifically, WiG [133] utilizes the hand-crafted statistical features while WiGeR [134] employs the signal matching approach. Moreover, we enhance the method WiGeR, by replacing the dynamic time warping (DTW) with a Gaussian kernel cross-correlator in the most recent work AAI18 [135], which constitutes a better baseline for our method. Also, to justify whether the CNN-LSTM features outperform normal deep features, we also conducted experiments on advanced CNN-based models (including modified AlexNet or ResNet for CSI data). These competitors can help prove the effectiveness of the CNN-LSTM features.

As each gesture has 40 samples for every person, we employ half as training and half as testing data from a total of $40 \times 10 \times 6$ samples. Two evaluation metrics, the overall accuracy and confusion matrix, are adopted. Each row of the confusion matrix indicates the predicted gesture class while each column represents the ground truth. As shown in Table 4.1, we compare our method with three gesture recognition methods and three state-of-the-art deep networks. Through cross-validation, our approach achieves the best average test accuracy of 92.8%, which improves the overall accuracy by 17.0%, 11.3%, 9.5% over WiG, WiGeR and AAI18, and 5.8%, 4.2%, 3.3% and 2.8% over modified AlexNet, ResNet18, ResNet50 and ResNet152,

respectively. WiG performs worst because they only use the SVM classifier with statistical features, which cannot generalize well. WiGeR and AAAI18 utilize the temporal signal matching techniques but they ignore the spatial features. On the contrary, classic CNNs do not make full use of temporal dynamics limited by the kernel size of convolution.

As shown in Figure 4.6, the first five confusion matrices give the detailed classification results. It is observed that our CNN-LSTM model achieves the best performance, followed by modified ResNet18. We find that *upward* and *pull* are the most misclassified gestures, since they are quite similar in terms of hand motions. In Figure 4.6(b) and 4.6(c), we observe quite similar confusion matrices because they are all based on signal matching. However, they are severely influenced by the intrinsic noise, which is partly overcome by the deep learning networks in Figure 4.6(d) and 4.6(e).

One-shot performance: For the number of training samples, we achieve one-shot learning using the Siamese architecture as well as a transferable pairwise loss. Under the same circumstance, we use gesture data from eight persons ($8 \times 6 \times 40$ samples) to train the Siamese model. Then the model takes 6 gesture inputs as training template and other $2 \times 6 \times 39$ samples as testing data. We compare the Euclidean distances between the testing sample and six templates, and the sample belongs to the class of its shortest neighbor. Through cross-validation, the one-shot accuracy reaches 89.5% shown as **SN-MMD** in Table 4.1. We also provide the confusion matrix of our approach in Figure 4.6(f), which is akin to CNN-LSTM. Since previous works never consider the issue of transferability and conventional deep networks cannot employ only one training sample, our approach is the first CSI-based gesture recognition system with high availability and satisfactory performance in practice.

Moreover, to fully understand the metric learning in our network, we change the output dimension of the fully connected layer to 2, and visualize the t-SNE embeddings of the features of 6×19 test gestures depicted in Figure 4.7. We observe that six clusters are generated in an unsupervised manner, and if we know the label for one sample of a cluster, it is reasonable to use the *nearest neighbor* algorithm to achieve one-shot learning. We can also observe that there are some black points (*Pull*) wrongly gathered around the cluster center of gray points (*Leftward*). This explains why *Pull* owns the worst accuracy of 73.7%, highly affected by *Leftwards*.

FIGURE 4.7: Test sample clustering results.

Meanwhile, these errors are caused by negative semantic transfer, and it may be further enhanced in the future.

The performance of dealing with environmental dynamics and individual heterogeneity: To test whether our cross-domain loss and metric learning mechanism contribute to addressing the environmental dynamics and individual heterogeneity, we conduct experiments by splitting the dataset according to environment or person. For the environmental dynamics, we use two domain data from a small conference room (S) and a large conference room (L). All the data in one environment and 20% data in the new environment are used for training while other 80% data is used for testing. Then we have two evaluation modes: $S \rightarrow L$ and $L \rightarrow S$. Shown in Table 4.2, we find that for traditional methods, using training data from two environments leads to confusion because they do not consider the environmental dynamics. Whereas, for our method, the features learned from the source domain are used in the target domain. The cross-domain loss further improves the robustness of the features and thus our method outperforms others by over 30%.

For the individual heterogeneity, we evaluate all methods from the same environment for controlled trial, and then select 5 persons group of CSI data as A and others as B. We still employ the data of one group as training ($5 \times 6 \times 40$) and another group as testing. Two tasks are formulated as $A \rightarrow B$ and $B \rightarrow A$. The results are reported by splitting the dataset for 5 times following our protocol. As shown in Table 4.2, the accuracies of WiG, WiGeR and AAAI-18 are about 45% accuracy, while our method achieves state-of-the-art performance of 85.7% and 83.5% for two transfer tasks. Compared with the baseline deep learning network (CNN-LSTM),

TABLE 4.2: Average Accuracy of Cross-domain Evaluation

Method	S→L	L→S	A→B	B→A
WiG	37.5	38.3	45.5	43.7
WiGeR	34.9	40.7	48.3	45.3
AAAI-18	35.3	42.3	49.1	46.7
CNN-LSTM	46.3	48.7	53.1	53.1
Ours	65.3	75.3	85.7	83.5

our architecture improves it by about 20%. Compared with the environmental dynamics, the heterogeneity problem is less severe in the gesture recognition task, but it still brings about 30% accuracy drop for other methods. Our method only drops by 7-9%, which demonstrates that our cross-domain loss and semantic metric learning effectively alleviate the problem of environmental dynamics and individual heterogeneity in a WiFi-based gesture recognition system.

HRR	95.0	5.0	0.0	0.0
HRL	2.5	97.5	0.0	0.0
FRR	0.0	0.0	55.0	45.0
FRL	0.0	0.0	52.5	47.5
	HRR	HRL	FRR	FRL

FIGURE 4.8: Confusion matrix of granularity test. We can see that hand-level gestures are easily classified while finger-level ones are tough to deal with due to the granularity of CSI data.

How can we increase the number of recognizable gestures: In practice, our method can involve more gestures if more types of gestures are available for training. Naturally, we should determine what level of human gestures can be captured by WiFi signals. We use a four-class classification experiment to demonstrate that its granularity is enough for recognizing hand-level motions rather than finger-level

ones. We choose four gestures including rolling left/right by hand (HRR/HRL) and the same gestures but only by one finger (FRR/FRL). We collect 100 samples of each gesture by one volunteer. We use 60 samples as training and 40 samples as validation. The proposed method added by a Softmax layer is adopted as classifier. The validation result is shown in Figure 4.8.

For the hand-level gestures, our method can achieve reasonable results, while for some subtle gesture, they cannot be classified due to the limitation of granularity. We find that the finger motions cannot affect CSI clearly, and thus no pattern is extracted. This investigation helps to identify proper gestures for recognition by WiFi-based systems.

4.5 Conclusion

This chapter establishes a gesture recognition system that allows pervasive implementation and achieves high accuracy. To avoid the use of dedicated infrastructure, our platform collects fine-grained CSI measurements from our developed WIFI-based IoT devices. The recognition task is then processed by a novel deep Siamese representation learning architecture. Technically, the proposed learning framework enriches the traditional Siamese network structure by providing additional temporal-spatial pattern learning capacity. From another point of view, the CNN and RNN based deep temporal-spatial feature learning is amended by the idea of Siamese to remove structured noise and to enable one-shot learning. Our algorithm possesses implementation convenience, high accuracy, and environmental transferability in one platform for gesture recognition. Experiments on real-world deployments show that our method performs better as compared to state-of-the-art approaches on feature learning, and achieves 89.5% accuracy under one-shot conditions.

Part II

Domain Adaptation for Robust Recognition

Chapter 5

Efficient Domain Adaptation for Transfer Learning in the Edge

Domain adaptation improves the generalization of machine learning models when confronting similar data distributions under new circumstances. However, existing domain adaptation approaches are designed for powerful servers with considerable computational resources. In this chapter, we formulate a new problem, Edge Domain Adaptation (EDA), to enhance the deployment of trained-once models in real-world edge devices. Section 5.1 introduces the background and real-world scenarios that motivate this research. Section 5.2 illustrates the details of MobileDA algorithm. In the Section 5.3, MobileDA is evaluated on both visual domain adaptation benchmark and our WiFi-based gesture recognition benchmark. Section 5.4 concludes the chapter.

5.1 Introduction

Consider a batch of household robots leveraging deep neural networks to recognize masses of common objects. The recognition model, trained using a large scale of standard datasets, is designed to be wider and deeper for better generalization. As robots are shipped to innumerable homes and environments, their recognition ability may deteriorate due to the issue of domain shift [61]. Besides, deploying such cumbersome model on the edge device puts forward much more stringent requirement on computational resources. It is well known that given sufficient

FIGURE 5.1: We transfer the novel knowledge from a sophisticated teacher model trained in the server to a simple student model running in the edge side while considering the domain shift in the new environment such as smart homes.

labeled data in a new environment, these two problems can be addressed by fine-tuning [136] and model compression [137]. However, it is unrealistic to annotate thousands of examples by the new owner of the robot and the model compression cannot be directly achieved by limited data with domain shift. How to address both problems simultaneously under constrained circumstances is quite challenging for now.

In the new environment, though labeled examples are difficult to obtain, robots can capture many images of objects randomly. These unlabeled examples share similar categories with the original training dataset but own an intractable domain shift. Utilizing the original labeled (source) data and new unlabeled (target) data, deep domain adaptation methods [61] can address the domain shift to some extent. There are also some methods such as model distillation [138] that enhance a simple model by very deep and wide models.

In this chapter, we first propose this research problem as Edge Domain Adaptation (EDA) problem. As shown in Fig. 5.1, its objective is to train a simple deep model that simultaneously addresses the domain shift problem for edge computing in IoT devices. Specifically, in the training phase, any novel deep models or advanced training techniques can be employed to learn better representation and conduct domain adaptation, but in the testing phase usually conducted in the edge, only a simple model is allowed. To this end, we propose an algorithm that effectively

FIGURE 5.2: Illustration of the proposed approach.

trains a simple student model that adapts to the new environment by diminishing domain disparity [61], and at the same time we utilize novel teacher models trained in the server to amend the student model by cross-domain knowledge distillation [138].

For domain adaptation, recent deep transfer learning methods mainly rely on minimizing the covariate shift between training and testing data during supervised learning, which can be achieved by the minimization of statistical measure of domain shift [61] and adversarial training [70]. As discussed in Chapter 2, the first category exploits various statistical measures to minimize the distribution divergence between two domains. Many metrics to measure domain difference, such as maximum mean discrepancy (MMD) [55] and covariance matrix [61], are proposed to obtain domain-invariant representations. These approaches accomplish adaptation in an end-to-end manner, and thus are quite easy to train. The adversarial-based methods stem from generative adversarial network (GAN) [63], tackling the adaptation problem by a minimax game between a domain classifier and the feature extractor [70]. They usually attain superior performance once the hyper-parameters are suitable for adversarial training. Nevertheless, the selection of apt settings still hinders its application. Note that these approaches cannot effectively reduce the distribution difference. In our scenario, since it is not possible to manually search hyper-parameters in the edge devices and the labels are mandatory for model compression, our method employs both the pseudo labels and the statistical way to diminish the domain disparity and obtain a novel target classifier.

As to model compression, various solutions have been developed such as network pruning [52], quantization [139] and distillation [138]. From the perspective of domain adaptation, it is observed that deep models are more effective than the adaptation methods based on shallow models, due to the better representation learning capacity of deep models. For example, for Office-31 dataset, the performance of an end-to-end adaptation method [140], based on AlexNet, cannot surpass that of source-only ResNet-50 model [141]. This inspires us to consider knowledge distillation which is more effective for the EDA problem. Through cross-domain distillation, not only is a simpler student model obtained, but also the domain adaptation is enhanced by the teacher network.

Our approach, illustrated in Fig. 5.2, pre-trains a teacher network in the server, and leverages its better generalization ability to teach the student model in the edge. We propose a unique loss function to train our model with three components: (1) cross-entropy loss on source labeled data, which ensures that source samples are discriminative; (2) distillation loss on target unlabeled data, which transfers more powerful knowledge from the teacher network; (3) domain confusion loss between training and testing domains, which seeks to learn transferable features across domains. In this way, the domain adaptation is amended by knowledge distillation, and on the other hand it helps the distillation process learn a domain-invariant representation. In the experiments, we evaluate MobileDA on a WiFi-based gesture recognition scenario, and three classic visual datasets including Office-31, Image-CLEF and Office-Home. The results demonstrate that under different level of domain shift, our simple student model achieves state-of-the-art performance on most of transfer tasks. In summary, the contributions of this chapter are addressed as follows:

- Edge Domain Adaptation (EDA) problem is formulated for a very practical scenario in IoT field, caused by limited computational resources and domain shift between the training data and realistic data.
- A novel method, namely MobileDA, is developed to deal with EDA. MobileDA reduces the domain shift by statistical metrics and further enhances the deep model by an innovative cross-domain distillation.

- Extensive experiments are conducted on IoT-based and visual benchmarks, whose results demonstrate that our approach surpasses other advanced methods and are more computationally-effective.

5.2 Method

5.2.1 Problem Definition and Notation

In unsupervised domain adaptation, we consider data from *source* and *target* domains. The source data refers to accessible datasets with labels, $\mathcal{D}_s = \{x_s^i, y_s^i\}_{i=1}^{N_s}$. The target examples come from the new environment and are only captured without labels, denoted as $\mathcal{D}_t = \{x_t^i\}_{i=1}^{N_t}$. Let $\mathcal{P}(X_s)$ and $\mathcal{Q}(X_t)$ be the marginal distributions of $\mathcal{D}_s$ and $\mathcal{D}_t$, respectively. The assumption of the UDA problem is that $\mathcal{P}(X_s) \neq \mathcal{Q}(X_t)$ but $\mathcal{P}(Y_s|X_s) = \mathcal{Q}(Y_t|X_t)$, which indicates that the source and target domains share the same tasks (i.e. label space Y) but have different data distributions. The typical deep networks are trained to produce an image feature extractor $f(x; \theta_{repr})$ and a category classifier θ_C , and predict with composite function $\Phi_{\theta_{repr}, \theta_C} = \sigma(\theta_C f(x; \theta_{repr}))$, where σ is the softmax function. In our model shown in Fig. 5.2, we employ a teacher network $\Phi_{\theta_{repr}^T, \theta_C^T}$ to supervise our objective student network $\Phi_{\theta_{repr}^S, \theta_C^S}$. The ultimate network should learn transferable representations from two domains, and can predict the labels $\{y_t^i\}_{i=1}^{N_t}$ for the target data.

5.2.2 Pre-training Teacher on Source Domain

We first train a teacher network using the available source labeled data. The teacher model Φ^T , trained on the server, can be deep and wide to learn robust representations. It consists of a teacher encoder $f(x; \theta_{repr}^T)$ which is usually a convolutional neural network (CNN) and a category classifier θ_C^T which is usually a fully connected layer with a softmax function. Our classification objective is defined as the cross-entropy loss

$$\mathcal{L}_{CE} = -\mathbb{E}_{(x_s, y_s) \sim (X_s, Y_s)} \sum_k [\mathbb{I}[y_s = k] \log p_{(k)}^T], \quad (5.1)$$

where p is the softmax of the classifier activations, i.e. $p^T = \sigma(\theta_C^T f(x_s; \theta_{repr}^T))$. More often than not, the teacher network can learn superior representations for better generalization than the student network in the target domain.

5.2.3 Distillation: Target Knowledge Transfer

From the experiment in Section 5.3, we discover that in most cases, using ResNet-50 brings about more improvement than most of the adaptation framework based on AlexNet. Intuitively, if we manage to train shallow networks under the supervision of a deeper network, any UDA method could be further improved. To this end, we design a soft target loss to transfer novel knowledge, inspired by the recent model distillation technique [138]. Considering an example $x_t^i \in \mathcal{D}_t$, the teacher network can produce class probabilities by a softmax layer that converts its logits $z^i = \{z_1^i, \dots, z_K^i\}$ to an output of probability q^i . In model distillation, a temperature T is incorporated into the standard softmax function as follows:

$$q_j^i = \frac{\exp(z_j^i/T)}{\sum_j \exp(z_j^i/T)}, \quad (5.2)$$

where T controls the attention of matching logits rather than one-hot labels. At higher temperature, the model pays more attention to learning logits which indicate the relationship between multiple categories. In MobileDA, we aim to employ the more powerful ability of the teacher network on the target domain. However, the target labels are not available. Motivated by the self-training, we iteratively select the samples from the target domain whose confidence score of the predicted category is higher than a threshold γ . Using the predicted category as the pseudo-label, we only transfer this part of ability to the student network, and the noises can be reduced via self-training. Suppose that we can obtain the predicted soft logits s_t of these high-confidence data on target domain using the teacher network

$$s_t = \sigma(\theta_C^T f(x_t; \theta_{repr}^T)/T). \quad (5.3)$$

Then the decision made by the teacher network $\Phi_{\theta_{repr}^T, \theta_C^T}$ can supervise the training of the student network on the target domain. The soft target loss is defined as

$$\mathcal{L}_{soft} = -\mathbb{E}_{(x_t, y_t) \sim (X_t, Y_t)} \sum_k [l(s_t, \theta_C^S f(x_t; \theta_{repr}^S))], \quad (5.4)$$

where l is the cost function. The target knowledge is transferred from $\Phi_{\theta_{repr}^T, \theta_C^T}$ to $\Phi_{\theta_{repr}^S, \theta_C^S}$ using the soft target loss. We can reason about the benefits of distillation in two-folds. Firstly, the student network learns more fine-grained representations from the teacher network. More importantly for the UDA problem, the student model receives noisy but self-updated supervisions in the target domain.

It is reasonable to envisage that overfitting may happen in $\mathcal{D}_t$. To avoid it, we also optimize the student network on the source domain leveraging hard labels as follows:

$$\mathcal{L}_{hard} = -\mathbb{E}_{(x_s, y_s) \sim (X_s, Y_s)} \sum_k [\mathbb{I}[y_s = k] \log p_{(k)}^S], \quad (5.5)$$

where $p^S = \sigma(\theta_C^S f(x_s; \theta_{repr}^S))$. The hard label loss ensures that the model still learns the discriminative features by adequate labeled source datasets. The joint learning of soft and hard label losses effectively prevents the overfitting to source or target distributions, accomplishing a trade-off between the two domains.

5.2.4 Adaptation: Diminishing Domain Disparity

Since the objective network obtains the supervisions of two domains separately, the last information to be considered is the distribution difference between two domains. In a typical deep network $\Phi_{\theta_{repr}, \theta_C}$, the convolution layers have been shown to be more generic while the fully connected (fc) layers are more task-specific. Therefore, we mainly modify the parameters of fc and slightly adjust the image encoder, which can be achieved by different learning rates. In our approach, we aim to reduce the domain disparity by minimization of the statistical distribution measure of two distinct domains. Plenty of previous works develop several measures including multi-kernel Maximum Mean Discrepancy (MK-MMD) [57], domain adversarial loss [142] and covariance matrix [61]. As the first two methods are highly sensitive to the hyper-parameters and their computation overloads have repercussion for edge latency, we introduce the CORAL loss [61] purely based on the distance between the covariance matrices in two domains.

In MobileDA, the domain confusion loss is defined as the distance between the

second-order statistics (covariance matrices) of the source and target representations produced by the student network $\Phi_{\theta_{repr}^S, \theta_C^S}$:

$$\mathcal{L}_{DC} = \frac{1}{4d^2} \|C_s - C_t\|_F^2, \quad (5.6)$$

where d is the number of samples or the batch size in the training phase, $\|\cdot\|_F^2$ denotes the Frobenius norm, and the covariance matrix of the source and target data (i.e. C_s, C_t , respectively) are given by:

$$C_s = \frac{1}{N_s - 1} (X_s^\top X_s - \frac{1}{X_s} (\mathbf{1}^\top X_s)^\top (\mathbf{1}^\top X_s)), \quad (5.7)$$

$$C_t = \frac{1}{N_t - 1} (X_t^\top X_t - \frac{1}{X_t} (\mathbf{1}^\top X_t)^\top (\mathbf{1}^\top X_t)), \quad (5.8)$$

where $\mathbf{1}$ is the column vector of appropriate dimension with all entries equal to one.

In summary, MobileDA is an end-to-end model formulated as an optimization problem of the joint loss function:

$$\begin{aligned} \min_{\theta_{repr}^S, \theta_C^S} \mathcal{L}(x_s, y_s, x_t; \theta_{repr}^T, \theta_C^T, \theta_{repr}^S, \theta_C^S) = \\ \mathcal{L}_{hard}(x_s, y_s; \theta_{repr}^S, \theta_C^S) + \lambda \mathcal{L}_{soft}(x_t; \theta_{repr}^T, \theta_C^T, \theta_{repr}^S, \theta_C^S) + \mu \mathcal{L}_{DC}(x_s, x_t; \theta_{repr}^S, \theta_C^S), \end{aligned} \quad (5.9)$$

where the hyper-parameters λ and μ determine the strength of distillation and domain confusion, respectively. These hyper-parameters can be obtained empirically. Then the loss function can be simply optimized by stochastic gradient descent and back-propagation. The optimization process is quite stable, which will be detailed in the experiment. After the optimization phase, the student model is obtained and can be employed in the embedded IoT devices, which is shown in Algorithm 3.

Algorithm 3 Mobile Domain Adaptation

Pretrain: Train the teacher model Φ^T in the server**for** each epoch **do** Obtain the high-confidence samples x_{hc} from x_t Calculate the soft logits s_t in Eq.(5.3) Calculate the soft target loss $\mathcal{L}_{soft}$ Calculate the CORAL loss $\mathcal{L}_{DC}$

Optimize the total loss in Eq.(5.9)

end for**Output:** the student model Φ^S in the IoT device

5.3 Experimental Results

In this section, we conduct extensive experiments to evaluate our MobileDA on the visual recognition task using three object recognition datasets that have distinct level of domain shift, and a WiFi-based gesture recognition dataset for non-visual adaptation using IoT sensors. We train the model using the source domain data and part of the target domain data, and test the model on the remaining part of the target domain data. The model can be transferred across devices with multiple settings of computational capacity, which simulates the training phase in the server and the testing phase in the IoT devices. We first evaluate the overall performance of MobileDA. Then, an ablation study is provided to explore the effectiveness of model distillation and different domain confusion methods in our framework. Apart from the accuracy of MobileDA, we further conduct experiment on an IoT embedded board with NVIDIA Jetson TX2 [143] to compare the computation latency of normal UDA model with ours.

5.3.1 Datasets

We employ the **WiFi Gesture Recognition Dataset** to evaluate the performance of MobileDA for spatial shift. Then, **Office-31** [94], **ImageCLEF-DA** and **Office-Home** are used for evaluation on visual domain adaptation. The datasets have been introduced in Chapter 2.

5.3.2 Setup and Implementation Details

We compare our methods with state-of-the-art deep transfer learning frameworks including both end-to-end and adversarial methods: Transfer Component Analysis (**TCA**) [69], Geodesic Flow Kernel (**GFK**) [144], Deep CORAL (**D-CORAL**) [61], Deep Adaptation Network (**DAN**) [3], Deep Reconstruction-classification Network (**DRCN**) [89], Joint Adaptation Network (**JAN**) [140], Domain Adversarial Neural Network (**DANN**) [93], Adversarial Discriminative Domain Adaptation (**ADDA**) [70], Consensus Adversarial Domain Adaptation (**CADA**) and Conditional Adversarial Domain Adaptation (**CDAN**) [145]. The overall evaluation protocol is different from the traditional one [3] that uses all the unlabeled data in the target domain for training and testing. In our case, we use 80% target unlabeled data for training, and all the target data for testing. Our protocol is more difficult because we use less data for training compared to the standard protocol. The reason why we design the particular protocol is to demonstrate that MobileDA indeed learns better representations from the teacher model, rather than only mimicking the softmax activation. We compare the average classification accuracy of each transfer task across three runs. For WiFi-based recognition, we reduplicated other methods based on the same network using their official hyper-parameters and implementations. For visual benchmarks, as the testing splits for two protocols are same, the results for comparison in Table 5.2, 5.3, 5.4 are directly derived from the original paper [145] based on the same network architecture.

We implement MobileDA based on **PyTorch** and employ the ImageNet pre-trained parameters as the encoder initialization. We choose AlexNet as our objective student model and ResNet-50 as the teacher model for overall comparison. The whole network is trained with the learning rate of η for the classifier and $\frac{\eta}{10}$ for the encoder. We adopt mini-batch stochastic gradient descent (SGD) with momentum of 0.9 and follow the strategy of adjusting learning rate in DANN [93]: $\eta_p = \frac{\eta_0}{(1+\alpha p)^\beta}$, where p is changing linearly from 0 to 1, and $\eta_0 = 0.01, \alpha = 10, \beta = 0.75$ are obtained by the importance-weighted cross-validation on Office-31. We set the temperature $T = 2$ and the threshold $\gamma = 0.7$. The strength of distillation λ is set to 0.9 and the strength of domain confusion μ is degraded as the same in [61]. We find that using these settings for all tasks bring about robust improvement, which is convenient to use for the situation when the model is trained on the IoT devices without human supervision.

In all visual benchmarks, we choose AlexNet [146] as the baseline model that is the same as our student model. For the WiFi dataset, we employ the SVHNNet as the teacher model [2] and the standard LeNet as the student model [1]. For MMD-based methods (DAN, RTN and JAN) and our ablation study, we use Gaussian kernel with bandwidth set to the median pairwise squared distances on the training data [57]. Any measure of domain difference is obtained using the deep representations output of layer *fc7* in AlexNet or *pool5* in ResNet-50.

5.3.3 Results

As shown in Table 5.1, due to the environment changes, the LeNet trained in the large room only performs 58.4% accuracy in the small room. Our approach achieves 31.8% improvement in the target domain (i.e. small room), and outperforms all other methods. The results demonstrate that our approach effectively reduces the effects of the environmental changes, leading to significant improvement in the new environment. As MobileDA requires less computation overhead, it is more suitable than the traditional domain adaptation approaches based on complicated models in the IoT side.

For various visual adaptation benchmarks, we further validate the performance of MobileDA from more perspectives. As shown in Table 5.2, the classification accuracies of MobileDA on *Office-31* dataset outperform existing methods on most of transfer tasks. We can see that the MobileDA model significantly promotes the accuracies of hard transfer tasks. For example, the accuracies of $\mathbf{A} \rightarrow \mathbf{D}$ and $\mathbf{W} \rightarrow \mathbf{A}$

TABLE 5.1: Comparisons on WiFi Gesture Recognition.

Source Target	Large Room Small Room
Method	Accuracy (%)
<i>Source-only</i>	58.4±0.7
DAN [3]	61.2±0.5
DRCN [89]	69.3±0.3
DANN [2]	68.2±0.2
ADDA [70]	71.5±0.3
CADA [97]	88.8±0.1
MobileDA	90.2±0.2

are improved by 27.9% and 26% respectively, compared to source-only AlexNet. The average accuracy even surpasses the performance of adversarial methods, e.g. ADDA and CDAN. Using only a baseline model, the adversarial methods [70] are shown to outperform statistical UDA methods [3], but the model distillation in MobileDA assists the student model to learn better representations and achieve excellent results.

TABLE 5.2: Accuracy (mean) of unsupervised domain adaptation on *Office-31* datasets across 3 independent runs.

Method	A→W	D→W	W→D	A→D	D→A	W→A	Avg
AlexNet [146]	45.2	88.1	96.2	47.4	38.2	36.1	58.5
TCA [69]	61.0	93.2	95.2	60.8	51.6	50.9	68.8
GFK [144]	60.4	95.6	95.0	60.6	52.4	48.1	68.7
D-CORAL [61]	61.6	95.7	99.2	66.8	52.8	51.5	72.1
DAN [3]	68.5	96.0	99.0	67.0	54.0	53.1	72.9
DANN [2]	73.0	96.4	99.2	72.3	53.4	51.2	74.3
ADDA [70]	73.5	96.2	98.8	71.6	54.6	53.5	74.7
JAN [140]	74.9	96.6	99.5	71.8	58.3	55.0	76.0
CDAN [145]	77.9	96.9	100.0	74.6	55.1	57.5	77.0
ResNet-50 [141]	68.3	96.7	99.2	69.1	61.7	60.0	75.8
MobileDA	71.5	97.4	99.8	75.3	63.4	62.1	78.3

However, the *Office-31* dataset is rather imbalanced, so we explore its performance on a neat dataset, *ImageCLEF-DA*. The results are reported in Table 5.3. It can be observed that our MobileDA model substantially outperforms other methods, while previous methods report minimal performance improvements room [145]. This is reasonable since the teacher model can achieve better generalization when the three domains in the *ImageCLEF-DA* dataset have equal size and balanced category, which magnifies the effect of model distillation for domain adaptation. We can observe that the ResNet-50 (teacher) model already has a similar average accuracy of 80.7% compared with the best adaptation accuracy of 80.9%. These discoveries further demonstrate the advantages of cross-domain adaptation.

TABLE 5.3: Accuracy (mean) of unsupervised domain adaptation on *ImageCLEF-DA* datasets across 3 independent runs.

Method	I→P	P→I	I→C	C→I	C→P	P→C	Avg
AlexNet [146]	62.0	70.8	82.7	66.8	50.8	80.5	68.9
DAN [3]	67.3	80.5	87.7	76.0	61.6	88.4	76.9
DANN [2]	66.5	81.8	89.0	79.8	63.5	88.7	78.2
JAN [140]	67.2	82.8	91.3	80.0	63.5	91.0	79.3
CDAN [145]	68.7	84.8	92.4	81.3	65.5	92.5	80.9
ResNet-50 [141]	74.8	83.9	91.5	78.0	65.5	91.2	80.7
MobileDA	75.5	89.5	94.7	87.7	71.8	95.3	85.8

TABLE 5.4: Accuracy (mean) of unsupervised domain adaptation on *Office-Home* datasets across 3 independent runs.

Method	Ar→Cl	Ar→Pr	Ar→Rw	Cl→Ar	Cl→Pr	Cl→Rw	Pr→Ar	Pr→Cl	Pr→Rw	Rw→Ar	Rw→Cl	Rw→Pr	Avg
AlexNet [146]	26.4	32.6	41.3	22.1	41.7	42.1	20.5	20.3	51.1	31.0	27.9	54.9	34.3
DAN [3]	31.7	43.2	55.1	33.8	48.6	50.8	30.1	35.1	57.7	44.6	39.3	63.7	44.5
DANN [2]	36.4	45.2	54.7	35.2	51.8	55.1	31.6	39.7	59.3	45.7	46.4	65.9	47.3
JAN [140]	35.5	46.1	57.7	36.4	53.3	54.5	33.4	40.3	60.1	45.9	47.4	67.9	48.2
CDAN [145]	38.0	48.7	61.7	37.3	53.6	57.6	32.9	44.1	62.1	48.2	49.5	70.7	50.4
ResNet-50 [141]	34.9	50.0	58.0	37.4	41.9	46.2	38.5	31.2	60.4	53.9	41.2	59.9	46.1
MobileDA	47.7	68.0	73.5	52.3	64.9	66.2	52.7	44.5	72.8	62.3	49.6	78.4	61.1

As to the most intractable dataset, *Office-Home*, the classification accuracies are reported in Table 5.4. Our approach substantially outstrips the comparison methods on all transfer tasks. We can observe that MobileDA surpasses the second-place approach by 10.7%. The reason is quite similar with the one explaining the large improvement on *ImageCLEF-DA*. *Office-Home* dataset has so many categories that the simple AlexNet can only obtain low in-domain classification accuracy. Incorporating a better representation from the teacher network, the baseline model obtains considerable improvement. This also indicates that only aligning the domain by minimizing the domain difference is far from enough, because the baseline model is limited even in the same domain.

It is desirable that MobileDA performs very well on both small and large domain shifts, and with balanced or imbalanced category. The overall results on three benchmarks show that by blending cross-domain distillation to traditional domain confusion, the proposed MobileDA allows a simple model to learn domain-invariant features with better knowledge from a deeper model.

TABLE 5.5: Ablation study on *ImageCLEF-DA*

Method	I→P	I→C	C→P	FLOPs ($\times 10^9$)
ResNet-18	72.8	91.2	65.2	1.8
MobileDA (AlexNet)	73.7	93.2	67.3	0.7
ResNet-50	74.8	91.5	65.5	3.8
MobileDA (AlexNet)	75.5	94.7	71.8	0.7
ResNet-101	76.3	93.5	70.7	7.6
MobileDA (AlexNet)	76.8	95.3	70.0	0.7
DAN (ResNet-50)	74.5	92.8	69.2	3.8
MobileDA (AlexNet)	75.5	94.7	71.8	0.7
MobileDA-Adversarial	73.5	91.7	70.2	-
MobileDA-MMD	75.2	94.0	71.5	-
MobileDA-CORAL	75.5	94.7	71.8	-

TABLE 5.6: Inference time on IoT devices

Method	FLOPs ($\times 10^9$)	Inference time (ms)
ResNet-18	1.8	26
ResNet-50	3.8	64
MobileDA (AlexNet)	0.7	12

FIGURE 5.3: Convergence and domain discrepancy. (a)(b) The loss and test error during training for Deep-CORAL and our MobileDA on $\mathbf{I} \rightarrow \mathbf{C}$ (*ImageCLEF-DA*). (c) $\mathcal{A}$ -distance on *Office-31* with transfer tasks of $\mathbf{A} \rightarrow \mathbf{W}$ and $\mathbf{W} \rightarrow \mathbf{D}$.

5.3.4 Ablation Study

In the ablation study, as we find that *Office-31* and *Office-Home* usually bring about larger standard deviations for multiple runs, we adopt *ImageCLEF-DA* as the experiment dataset. Three transfer tasks are considered at distinct level of difficulty. We investigate our MobileDA from three perspectives including different teacher networks, latency comparison and the effect of several domain shift measures.

Teacher network. To further study the impact of model distillation, we employ three novel teachers including ResNet-18, ResNet-50 and ResNet-101. Shown in Table 5.5, we draw the source-only errors of these networks compared with AlexNet under their supervisions. It is easy to see that the better the teacher is, the more improvement the student yields. Also, we can see that apart from distillation, the domain confusion loss contributes to the student’s learning with better representations and accuracy improvement of 1% to 5%.

Computation latency. Considering the real application of household robots, we aim to train a model that leads to better accuracy and also lower latency. We compare the Floating Point Operations (FLOPs) of our MobileDA network based

FIGURE 5.4: Confusion matrix of source-only AlexNet (70.8%), Deep-CORAL (73.3%) and MobileDA (89.5%) on the transfer task of $\mathbf{P} \rightarrow \mathbf{I}$ on *ImageCLEF-DA*.

FIGURE 5.5: t-SNE plots of MobileDA features on the transfer task of $\mathbf{P} \rightarrow \mathbf{I}$ on *ImageCLEF-DA*.

on AlexNet and DAN [3] based on ResNet-50 that achieve very similar classification accuracies on target domain. As shown in Table 5.5, we can see that our method achieves similar results with a lower computational load. When deploying these methods in the embedded GPU, the latency will be very different due to the FLOPs gap. We employ the embedded GPU, NVIDIA Jetson TX2, to measure the inference time of the model. The inference time refers to the forward duration of the model for one input sample. The results are shown in Table 5.6. The inference of the proposed MobileDA only takes 12ms for one sample while that of ResNet-50 takes 64ms with similar classification accuracy. This demonstrates that our model has better efficiency on IoT devices while preserving good performance.

Hyper-parameter sensitivity. We explore the sensitivity of hyper-parameters. As μ always degrades during training following the strategy in [61], we only investigate the weight of cross-domain knowledge distillation, *i.e.* λ . We conducted the experiment on ImageCLEF-DA ($\mathbf{I} \rightarrow \mathbf{P}$). As shown in Table 5.7, the performance

is improved as the weight increases. When the weight is larger than 0.6, the improvement gain is similar, which infers that the cross-domain distillation results in significant improvement and such improvement is rather robust when the weight λ is great enough.

TABLE 5.7: Accuracy on *ImageCLEF-DA* ($\mathbf{I} \rightarrow \mathbf{P}$) *w.r.t* λ .

λ	0	0.2	0.4	0.6	0.8	1.0
Accuracy (%)	66.5	68.2	70.8	74.5	75.2	75.3

Domain discrepancy measure. We explore the effect of different ways of reducing domain discrepancy for our method, including adversarial loss, MK-MMD and CORAL. Adversarial UDA approaches are very popular and its end-to-end adversarial training loss can be found in DANN [93]. MK-MMD is proposed in [57] to conduct *two-sample test* in the Reproducing Kernel Hilbert Space (RKHS) using a family of kernels. We compare two losses based on our approach using AlexNet (student) and ResNet-50 (teacher). The results are reported in Table 5.5. We can find that the adversarial way falls short of our expectations, and the reason may lie in its instability of training. Whereas, the MK-MMD based MobileDA achieves similar results as the CORAL based one. The difference of accuracies only varies from 0.3% to 0.7%. To get the best adaptation result, we choose CORAL as our metric to effectively diminish the domain discrepancy.

5.3.5 Analysis

Convergence. To further study the generalization of MobileDA on source and target domains, we draw the losses during training for Deep-CORAL [61] and our method. We conduct experiments on *ImageCLEF-DA* dataset with $\mathbf{I} \rightarrow \mathbf{C}$ transfer task. As depicted in Fig. 5.3(a), the losses of Deep-CORAL converge faster. The source and target accuracies reach the best of 100% and 82% in 5 epochs. Whereas, for MobileDA, the cross-domain distillation helps the student model to achieve better target accuracy of 94.7% within 10 epochs, which may avoid overfitting on the source domain to some extent. We can also see that with the help of distillation, the CORAL loss starts from a smaller value and can be further minimized compared with the original Deep-CORAL.

Distribution Discrepancy. The learning theory of domain adaptation [48] proposed $\mathcal{A}$ -distance as a measure of distribution discrepancy. The empirical $\mathcal{A}$ -distance is defined as $d_{\mathcal{A}} = 2(1 - 2\epsilon)$, where ϵ is the generalization error of a classifier trained to discriminate the source and target domains. We employ a kernel SVM as the binary classifier and compare MobileDA with source-only AlexNet and DANN [93]. The classifier is trained using the features of three methods on tasks $\mathbf{A} \rightarrow \mathbf{W}$ as well as $\mathbf{W} \rightarrow \mathbf{D}$. As described in Fig. 5.3(c), we can observe that $d_{\mathcal{A}}$ on MobileDA features are much smaller than the basic AlexNet, so the domain shift between source and target features are substantially reduced. Nevertheless, we find that the $\mathcal{A}$ -distances of MobileDA and DANN are quite similar, but our MobileDA achieves better classification error on the target domain. This indicates that when the distribution difference has been small and similar, cross-domain distillation transfers better in-class knowledge to our baseline model. The phenomenon implies that only maximizing domain confusion is not enough, and it motivates us to develop model distillation for UDA to learn a more discriminative representation.

Confusion Matrix. Apart from the comparison of overall accuracies, we also investigate the category relationship by the confusion matrix. In Fig. 8.4, we compare our approach with source-only AlexNet and Deep-CORAL on $\mathbf{P} \rightarrow \mathbf{I}$ using *ImageCLEF-DA* dataset. We can see that by aligning domains and learning better representations, the proposed MobileDA improves the source-only model by 18.7%. We can also see that the domain alignment slightly boosts the classification, but from the confusion matrices of AlexNet and Deep-CORAL, we find that the pure alignment may result in some faults such as the second and the last rows in Fig. 5.4(b). This observation also suggests us to learn a superior representation rather than only aligning domain difference.

Representation Visualization We visualize the network activations of task $\mathbf{P} \rightarrow \mathbf{I}$ (12 classes) by t-SNE [147] as shown in Fig. 5.5. For AlexNet and our method, we plot them using the domain label shown in the left and category label shown in the right. Before adaptation, we can see that the source and target domains are not aligned but 12 categories are intangibly classified. After adaptation, the features of the source and target domains are better aligned while most categories are perfectly discriminated with obvious borders. This is resulted from the help

of cross-domain distillation that transfers better semantic representations to the student target model.

5.4 Conclusion

In this chapter, we study the edge domain adaptation problem that aims to learn a model considering domain-shifted data under rather limited conditions. To deal with the problem, we have proposed MobileDA model by cross-domain distillation. MobileDA diminishes the distribution difference as most UDA methods while simultaneously learns a superior representation from a deeper model with better generalization. With the supervision of cross-domain distillation, any UDA approach can be enhanced. Extensive experiments demonstrated the efficacy of MobileDA. It achieves state-of-the-art performance on four benchmarks and shows faster inference speeds than the models that achieve similar results on embedded IoT devices with GPU.

Chapter 6

Stable and Comprehensive Domain Adaptation: Max-margin Domain-Adversarial Training

For real-world domain adaptation solutions, a basic requirement is that the algorithm should easily reach convergence without manual human tuning. This chapter aims to stabilize adversarial domain adaptation by proposing Max-margin Domain-Adversarial Training (MDAT). The motivation and related works of MDAT are introduced in Section 6.1. In Section 6.2, the problem formulation of MDAT is elaborated and the optimal solution is discussed. Extensive experiments on visual and WiFi adaptation tasks demonstrate the training stability and the improved performance of MDAT in Section 6.3. The chapter is summarized in Section 6.4.

6.1 Introduction

Prevailing methods for unsupervised domain adaptation (UDA) are mostly based on *domain alignment* which aims to learn domain-invariant features by reducing the distribution discrepancy between the source and target domain using some pre-defined metrics such as maximum mean discrepancy [55, 57]. Recently, Ganin and Lempitsky [2] proposed to achieve domain alignment by domain-adversarial

training (DAT) that reverses the gradients of a domain classifier to maximize domain confusion. Having yielded remarkable performance gain, DAT was employed in many subsequent UDA methods [145, 148, 149].

Nevertheless, DAT with gradient reverse layers still face three critical restrictions when applying it to practical scenarios. (1) DAT cannot continuously provide effective gradients for learning domain-invariant representations. The reason is that the binary domain classifier has high-capacity to discriminate two domains and thus overwhelms adversarial training, which is usually solved by manually adjusting the weights of adversarial loss according to specific tasks such as [72]. (2) DAT cannot deal with pixel-level *domain shift* that is frequently encountered in visual tasks [66]. (3) The domain-invariant features learned by DAT are only based on intuition and learning theory [48] but difficult to interpret, which impedes the investigation of the underlying mechanism of adversarial domain adaptation.

To overcome the aforementioned difficulties, we propose a novel adversarial approach, namely Max-margin Domain-Adversarial Training (MDAT), to realize stable and comprehensive (*i.e.* both feature-level and pixel-level) domain alignment. MDAT works based on a carefully-designed Adversarial Reconstruction Network (ARN). Specifically, ARN consists of a shared feature extractor, a label predictor, and a reconstruction network (*i.e.* decoder) that serves as a domain classifier. MDAT enables an adversarial game between the feature extractor and the decoder. The decoder focuses on reconstructing features on source domain and pushing target features away by a *margin*, while the feature extractor aims to fool the decoder by generating target features that can be reconstructed. In this adversarial way, three critical issues are subtly solved: (1) the max-margin loss reduces the discriminative capacity of domain classifier, balancing and stabilizing domain-adversarial training; (2) without involving any new network structures, MDAT achieves both pixel-level and feature-level domain alignment; (3) reconstructing adapted features to images reveals how the source and target domains are aligned by adversarial training. We evaluate ARN with MDAT on both visual and non-visual UDA benchmarks. It shows more training procedure and achieves significant improvement to DAT on all tasks with pixel-level or higher-level *domain shift*. We also observe that it is insensitive to the choices of hyperparameters and as such is favorable for replication in practice. In principle, our approach is generic and can be

used to enhance any domain adaptation method that leverages domain alignment as an ingredient.

FIGURE 6.1: The proposed architecture is composed of a shared feature extractor G_e for two domains, a label predictor G_y and a reconstruction network G_r . In addition to the basic supervised learning in the source domain, our adversarial reconstruction training enables the extractor G_e to learn domain-invariant features. Specifically, the network G_r aims to reconstruct the source samples x^s and to impede the reconstruction of the target samples x^t , while the extractor G_e tries to fool the reconstruction network in order to reconstruct the target samples x^t .

6.2 Problem Formulation

6.2.1 Problem Definition and Notations

In unsupervised domain adaptation, we assume that a model works with a labeled dataset $\mathbf{X}_S$ and an unlabeled dataset $\mathbf{X}_T$. Let $\mathbf{X}_S = \{(\mathbf{x}_i^s, y_i^s)\}_{i \in [N_s]}$ denote the labeled dataset of N_s samples from the source domain, and the certain label y_i^s belongs to the label space Y that is a finite set ($Y = 1, 2, \dots, K$). The other dataset $\mathbf{X}_T = \{\mathbf{x}_i^t\}_{i \in [N_t]}$ has N_t samples from the target domain but has no labels. We further assume that two domains have different distributions, *i.e.* $\mathbf{x}_i^s \sim \mathcal{D}_S$ and $\mathbf{x}_i^t \sim \mathcal{D}_T$. In other words, there exists some *domain shift* [48] between $\mathcal{D}_S$ and $\mathcal{D}_T$. The ultimate goal is to learn a model that can predict the label y_i^t given the target input $\mathbf{x}_i^t$.

6.2.2 Unbalanced Minimax Game in Domain-Adversarial Training

To achieve domain alignment, Domain-Adversarial Training (DAT) is a minimax game between a shared feature extractor F for two domains and a domain classifier D . The domain classifier is trained to determine whether the input sample belongs to the source or the target domain while the feature extractor learns to deceive the domain classifier, which is formulated as:

$$\min_F \max_D \mathbb{E}_{x \sim D_s} [\ln F(x)] + \mathbb{E}_{x \sim D_t} [\ln (1 - D(F(x)))]. \quad (6.1)$$

We usually utilize Convolutional Neural Network (CNN) as the feature extractor and fully connected layers (FC) as the domain classifier. Theoretically, DAT reduces the cross-domain discrepancy and helps learn domain-invariant representations [93]. However, the training of DAT is rather unstable. Without sophisticated tuning of the hyper-parameters, DAT cannot often reach the convergence. Through empirical experiments, we observe that such instability is due to the imbalanced adversarial game between D and F . The binary domain discriminator D can easily achieve convergence with very high accuracy at an early training epoch, while it is much harder for the feature extractor F to fool the domain discriminator and to simultaneously perform well on the source domain. In this sense, there is a strong likelihood that the domain classifier overwhelms DAT, and the only solution is to palliate the training of D by tuning the hyper-parameters according to different tasks. In our method, we restrict the capacity of the domain classifier so as to form a minimax game in a harmonious manner. Inspired by the max-margin loss in Support Vector Machine (SVM) [150] (*i.e.* hinge loss), if we push the source domain and the target domain away by a margin rather than as far as possible, then the training task of F to fool D becomes much easier. For a binary domain classifier, we define the margin loss as

$$\mathcal{L}_{mg}(y) = [0, m - t \cdot y]^+, \quad (6.2)$$

where y is the predicted domain label, $[\cdot]^+$ is $\max(0, \cdot)$, m is a positive margin and t is the ground truth label for two domains (assuming $t = -1$ for the source domain and $t = 1$ for the target domain). Then we introduce our MDAT scheme based on an innovative network architecture.

6.2.3 Max-margin Domain-Adversarial Training

Besides the training instability issue, DAT also suffers from restrictive feature-level alignment – lack of pixel-level alignment. To realize stable and comprehensive domain alignment together, we first propose an Adversarial Reconstruction Network (ARN) and then elaborate MDAT.

As depicted in Figure 6.1, our model consists of three parts including a shared feature extractor G_e for both domains, a label predictor G_y and a reconstruction network G_r . Let the feature extractor $G_e(\mathbf{x}; \theta_e)$ be a function parameterized by θ_e which maps an input sample $\mathbf{x}$ to a deep embedding $\mathbf{z}$. Let the label predictor $G_y(\mathbf{z}; \theta_y)$ be a task-specific function parameterized by θ_y which maps an embedding $\mathbf{z}$ to a task-specific prediction $\hat{y}$. The reconstruction network $G_r(\mathbf{z}; \theta_r)$ is a decoding function parameterized by θ_r that maps an embedding $\mathbf{z}$ to its corresponding reconstruction $\hat{\mathbf{x}}$.

The first learning objective for the feature extractor G_e and the label predictor G_y is to perform well in the source domain. For a supervised K -way classification problem, it is simply achieved by minimizing the negative log-likelihood of the ground truth class for each sample:

$$\mathcal{L}_{task} = \sum_{i=1}^{N_s} \mathcal{L}_y(\mathbf{x}_i^s, \mathbf{y}_i^s) = - \sum_{i=1}^{N_s} \mathbf{y}_i^s \cdot \log G_y(G_e(\mathbf{x}_i^s; \theta_e); \theta_y), \quad (6.3)$$

where $\mathbf{y}_i^s$ is the one-hot encoding of the class label y_i^s and the logarithm operation is conducted on the softmax predictions of the model.

The second objective is to render feature learning to be domain-invariant. This is motivated by the *covariate shift* assumption [151] that indicates if the feature distributions $S(\mathbf{z}) = \{G_e(\mathbf{x}; \theta_e) | \mathbf{x} \sim \mathcal{D}_S\}$ and $T(\mathbf{z}) = \{G_e(\mathbf{x}; \theta_e) | \mathbf{x} \sim \mathcal{D}_T\}$ are similar, the source label predictor G_y can achieve a similar accuracy in the target domain. To this end, we first design a decoder network G_r that serves as a domain classifier. In MDAT, we train the decoder network G_r to only reconstruct the features in the source domain and to push the features in the target domain away by a margin. In this way, the decoder has the functionality of distinguishing the source domain from the target domain. The objective of training G_r is formulated

as

$$\min_{\theta_r} \sum_{i=1}^{N_s+N_t} \mathcal{L}_{mg}(\mathcal{L}_r(\mathbf{x}_i)) = \min_{\theta_r} \sum_{i=1}^{N_s} \mathcal{L}_r(\mathbf{x}_i^s) + \sum_{j=1}^{N_t} [m - \mathcal{L}_r(\mathbf{x}_j^t)]^+, \quad (6.4)$$

where m is a positive margin and $\mathcal{L}_r(\cdot)$ is the mean squared error (MSE) term for the reconstruction loss defined as

$$\mathcal{L}_r(\mathbf{x}) = \|G_r(G_e(\mathbf{x}; \theta_e); \theta_r) - \mathbf{x}\|_2^2, \quad (6.5)$$

where $\|\cdot\|_2^2$ denotes the squared L_2 -norm. Compared to the normal binary domain classifier (*e.g.* fully connected layers), the decoder network is tailored as a smoothing domain discriminator by keeping two domains by a specific margin rather than as far as possible.

On the other hand, to form an adversarial game, the feature extractor G_e learns to deceive G_r such that the learned target features are indistinguishable to the source ones, which is formulated as:

$$\min_{\theta_e} \sum_{j=1}^{N_t} \mathcal{L}_r(\mathbf{x}_j^t). \quad (6.6)$$

Then the whole learning procedure of ARN with MDAT can be formulated as:

$$\min_{\theta_e, \theta_c} \sum_{i=1}^{N_s} \mathcal{L}_y(\mathbf{x}_i^s, \mathbf{y}_i^s) + \alpha \sum_{j=1}^{N_t} \mathcal{L}_r(\mathbf{x}_j^t), \quad (6.7)$$

$$\min_{\theta_r} \sum_{i=1}^{N_s} \mathcal{L}_r(\mathbf{x}_i^s) + \sum_{j=1}^{N_t} [m - \mathcal{L}_r(\mathbf{x}_j^t)]^+, \quad (6.8)$$

where $\mathcal{L}_y$ defined in Eq.(6.3) is the negative log-likelihood of the ground truth class for labeled sample $(\mathbf{x}_i^s, \mathbf{y}_i^s)$ and α controls the interaction of the loss terms. In the following section, we derive an optimal solution of MDAT and provide theoretical justifications on how MDAT reduces the distribution discrepancy for UDA.

6.2.4 Optimal Solution of MDAT

Considering the adversarial game between a reconstruction network R and a feature extractor E (*i.e.* G_r and G_e in our network, respectively), we prove that if the feature extractor E maps both source domain $x^s \sim P_S(x^s)$ and target domain $x^t \sim$

$P_T(x^t)$ to a common feature space $\mathcal{Z} = \{z = E(x) | z \sim P_z(z)\}$, the MDAT system reaches a Nash Equilibrium. This theoretically explains how MDAT enables the feature extractor to learn domain-invariant features. Similar to EBGAN [152], we assume E and R have infinite capacity. Denote R as the MSE of the reconstruction network. We first define two objectives:

$$V(E, R) = \int_{x^s, x^t} \mathcal{L}_{mg}(x^s, x^t) P_S(x^s) P_T(x^t) dx^s dx^t \quad (6.9)$$

$$U(E, R) = \int_{x^t} \mathcal{L}_r(x^t) P_T(x^t) dx^t \quad (6.10)$$

In MDAT, we train the feature extractor E to minimize the quantity $V(E, R)$ and train the reconstruction network R to minimize the quantity $U(E, R)$. A Nash equilibrium of our system is a pair (E^*, R^*) that satisfies:

$$V(E^*, R^*) \leq V(E^*, R) \quad \forall R \quad (6.11)$$

$$U(E^*, R^*) \leq U(E, R^*) \quad \forall E \quad (6.12)$$

Theorem 6.2.1. If a feature extractor E maps both source domain $x^s \sim P_S(x^s)$ and target domain $x^t \sim P_T(x^t)$ to a common feature space $\mathcal{Z} = \{z = E(x) | z \sim P_z(z)\}$, the system reaches a Nash equilibrium (E^*, R^*) and $V(E^*, R^*) = m$.

Proof. We first prove Eq.(6.11):

$$\begin{aligned} V(E^*, R) &= \int_{x^s} P_S(x^s) R(E^*(x^s)) dx^s \\ &\quad + \int_{x^t} P_T(x^t) [m - R(E^*(x^t))]^+ dx^t \end{aligned} \quad (6.13)$$

$$= \int_z \left(P_z(z) R(z) + P_z(z) [m - R(z)]^+ \right) dz \quad (6.14)$$

As we know $f(x) = x + [m - x]^+$ is monotonically increasing on $[0, +\infty)$, $V(E^*, R)$ reaches its minimum when $R^*(z) = 0$:

$$V(E^*, R^*) = m \int_z P_z(z) dz = m \quad (6.15)$$

When $R^*(z) = 0$, we expand $U(E, R^*)$ in Eq.(6.12):

$$U(E, R^*) = \int_{x^t} R^*(E(x^t)) P_T(x^t) dx^t \quad (6.16)$$

$$= \int_z R^*(z) dz = 0 \quad (6.17)$$

As $U(E, R^*) \geq 0$, we get $U(E^*, R^*) \leq U(E, R^*)$.

It can be easily observed that the optimal solution of MDAT is a Nash equilibrium when the feature extractor maps two domains into a common feature space, *i.e.* aligning the distributions in the feature space.

6.2.5 Connection to Domain Adaptation Theories

We further investigate how MDAT connects the learning theory of domain adaptation. The rationale behind domain alignment is motivated from the learning theory of non-conservative domain adaptation problem [48]:

Theorem 6.2.2. Let $\mathcal{H}$ be the hypothesis space where $h \in \mathcal{H}$. Let $(\mathcal{D}_S, \epsilon_s)$ and $(\mathcal{D}_T, \epsilon_t)$ be the two domains and their corresponding generalization error functions. The expected error for the target domain is upper bounded by

$$\epsilon_t(h) \leq \epsilon_s(h) + \frac{1}{2} d_{\mathcal{H} \Delta \mathcal{H}}(\mathcal{D}_S, \mathcal{D}_T) + \lambda, \forall h \in \mathcal{H}, \quad (6.18)$$

where the ideal risk $\lambda = \min_h [\epsilon_s(h) + \epsilon_t(h)]$, and the distribution discrepancy $d_{\mathcal{H} \Delta \mathcal{H}}(\mathcal{D}_S, \mathcal{D}_T) = 2 \sup_{h, h' \in \mathcal{H}} |\mathbb{E}_{x \sim \mathcal{D}_S}[h(x) \neq h'(x)] - \mathbb{E}_{x \sim \mathcal{D}_T}[h(x) \neq h'(x)]|$

Theoretically, when we minimize the $\mathcal{H} \Delta \mathcal{H}$ -distance, the upper bound of the expected error for the target domain is reduced accordingly. Assuming a family of domain classifiers $\mathcal{H}_r$ to be rich enough to contain the symmetric difference hypothesis set of a family of label predictors $\mathcal{H}_p$, such that $\mathcal{H}_r = \mathcal{H}_p \Delta \mathcal{H}_p = \{\delta | \delta = \min(h \oplus h', m), h, h' \in \mathcal{H}_p, m \in [0, 1]\}$ where $\oplus$ is the XOR-function, the empirical $\mathcal{H}_p \Delta \mathcal{H}_p$ -distance has an upper bound *w.r.t.* the optimal domain classifier δ :

$$d_{\mathcal{H}_p \Delta \mathcal{H}_p}(\hat{\mathcal{D}}_S, \hat{\mathcal{D}}_T) \leq 2 \sup_{\delta \in \mathcal{H}_r} |\mathbb{E}_{\mathbf{z} \sim \hat{\mathcal{D}}_S}[\delta(\mathbf{z}) = 0] + \mathbb{E}_{\mathbf{z} \sim \hat{\mathcal{D}}_T}[\delta(\mathbf{z}) = m] - 1|, \quad (6.19)$$

where $\hat{\mathcal{D}}_S$ and $\hat{\mathcal{D}}_T$ denote the distributions of the source and target feature space $\mathcal{Z}_S$ and $\mathcal{Z}_T$, respectively. We rewrite the feature $\mathbf{z}$ as $E(\mathbf{x})$, and the reconstruction network R serves as the hypothesis δ . Then if we replace *sup* with *max* and minimize the term *w.r.t.* E , we obtain

$$\min_E \max_R |\mathbb{E}_{\mathbf{x} \sim \mathcal{D}_S}[R(E(\mathbf{x})) = 0] + \mathbb{E}_{\mathbf{x} \sim \mathcal{D}_T}[R(E(\mathbf{x})) = m] - 1|. \quad (6.20)$$

This equation is the objective of MDAT in Eq.(6.7) and Eq.(6.8) where R pushes two domains away by a margin m while E aims to minimize it, which also conforms with $V(E^*, R^*) = m$ at the Nash equilibrium.

6.2.6 Merits: MDAT vs. DAT

Compared with the conventional DAT with a binary domain classifier [2], the proposed MDAT is more attractive and incorporates new merits conceptually and theoretically:

(1) Stable convergence and effective gradients. MDAT restrains the overwhelming capacity of domain classifier and forms a balanced adversarial game, which tackles two problems caused by *large or small weight of gradient reverse in DAT*. (a) MDAT can easily reach convergence and thus enables unmanned UDA in practical scenarios. (b) The decoder network can continuously offer effective gradients to the feature extractor, leading to more domain-invariant representation.

(2) Cyclic constraint and better alignment. The decoder network introduces a cyclic constraint that helps *learn target-specific and pixel-level features by unsupervised reconstruction* [88, 90]. The asymmetric objectives for two domain in Eq.(6.7) further lead to better alignment as discussed in ADDA [70]. In comparison, DAT only utilizes a bit of information to align feature-level distributions.

(3) Better interpretability. MDAT can be well explained by Nash equilibrium and learning theories. More intuitively, MDAT enables to visualize the features by directly decoding features to images. In Table 6.6, the visualizations show that the decoded target images become source-like after MDAT, which implies that two domains are successfully mapped into a source-like common feature space.

6.3 Experiment

We evaluate the proposed approach on several visual and non-visual UDA tasks with varying degrees of *domain shift*. Then detailed analyses are conducted *w.r.t.* toy dataset, parameter sensitivity, gradient and feature visualization.

6.3.1 Setup

We firstly leverage **Digits** dataset for evaluation of low-level domain shift. Then **Office-Home** and **WiFi Gesture Recognition** are utilized to verify the effectiveness of MDAT on large domain shift and non-visual domain shift, respectively. The datasets have been introduced in Chapter 2.

We compare our method with state-of-the-art UDA methods that perform three ways of domain alignment. Specifically, **MMD** regularization [3] and **CORAL** [61] are based on statistical distribution matching. **DRCN** [89] and **DSN** [90] use the reconstruction network for UDA, while more prevailing UDA methods adopt domain-adversarial training including **DANN** [2], **ADDA** [70], **CyCADA** [66], **CDAN** [145], **MCD** [73], **CADA** [97], **TransNorm** [153] and **BSP** [148]. The baseline results are reported from their original papers where available.

We used **Pytorch** to implement our model. For **Digits** dataset, we follow the same protocol in [66] and the same network architecture of [2] which is shown in Table 6.1. For **Office-Home**, we adopt ResNet-50 pretrained on ImageNet as our backbone. According to the standard protocols in [3], we employ all the labeled source samples and unlabeled target samples for training. For **WiFi Gesture Recognition** data, we employ the modified LeNet and the standard protocol in [97]. The designs of G_r are the inverse of G_e with pooling operation replaced by upsampling. We fix $\alpha = 0.02$ and $m = 5$ in all the experiments, which are obtained on **SVHN**→**MNIST** by Bayesian optimization [154]. We adopt mini-batch SGD optimizer with momentum of 0.9 and the progressive training strategy in DANN [2].

Layer Index	Feature Extractor	Decoder Network	Label Predictor
0	$32 \times 32 \times 3$ Image		
1	5×5 conv. 64 ReLU	2048 dense, ReLU	10 dense, softmax
2	3×3 max-pool, stride 2	3072 dense, ReLU	
3	5×5 conv. 64 ReLU	5×5 conv. 128 ReLU	
4	3×3 max-pool, stride 2	upsample 2	
5	5×5 conv. 128 ReLU	5×5 conv. 64 ReLU	
6	3072 dense, dropout, ReLU	upsample 2	
7	2048 dense, dropout ReLU	5×5 conv. 64 ReLU	

TABLE 6.1: The network architecture used in the *Digits* experiments.

6.3.2 Overall Results

The classification accuracies on Digits are shown in Table 6.2. Our method outperforms all other methods on four transfer tasks. Specifically, for **SVHN**→**MNIST** where severe pixel-level domain shift exists, our method significantly improves **DANN** by 22.7%, which justifies the efficacy of **ARN** for addressing pixel-level shift. Our method also performs well when the target domain is quite small, achieving 98.6% accuracy on **MNIST**→**USPS**. In Table 6.3, our method improves the source-only model by 32.9% on the WiFi spatial adaptation problem, which indicates that **MDAT** is also helpful for non-visual domain adaptation problem. Table 6.4 shows the performance on large-scale dataset and **MDAT** yields better performance against other domain alignment approaches.

FIGURE 6.2: The *inter-twining moons* toy problem. Red and green dots indicate source samples, while blue dots are target samples. Black lines indicate the changes of decision boundaries during 10 epochs training.

TABLE 6.2: We compare with general, statistics-based (**S**), reconstruction-based (**R**) and adversarial-based (**A**) domain adaptation approaches. We repeated each experiment for 3 times and report the average and standard deviation (std) of the test accuracy in the target domain.

	Source Target	MNIST USPS	USPS MNIST	SVHN MNIST	SYN SVHN
<i>Source-Only model</i>		78.2	63.4	54.9	86.7
<i>Train on target</i>		96.5	99.4	99.4	91.3
[S] MMD [3]		81.1	-	71.1	88.0
[S] CORAL [61]		80.7	-	63.1	85.2
[R] DRCN [89]		91.8	73.7	82.0	87.5
[R] DSN [90]		91.3	-	82.7	91.2
[A] DANN (DAT) [93]		85.1	73.0	74.7	90.3
[A] ADDA [70]		89.4	90.1	76.0	-
[A] CDAN [145]		93.9	96.9	88.5	-
[A] CyCADA [66]		95.6	96.5	90.4	-
[A] BSP+DANN [148]		94.5	97.7	89.4	-
[A] MCD [73]		96.5	94.1	96.2	-
[A] CADA [97]		96.4	97.0	90.9	-
ARN w.o. MDAT		93.1±0.3	76.5±1.2	67.4±0.9	86.8±0.5
ARN with MDAT (proposed)		98.6±0.3	98.4±0.1	97.4±0.3	92.0±0.2

TABLE 6.3: Comparisons on WiFi gesture recognition.

Method	Accuracy (%)
<i>Source-only</i>	58.4±0.7
[S] MMD [3]	61.2±0.5
[R] DRCN [89]	69.3±0.3
[A] DANN (DAT) [2]	68.2±0.2
[A] ADDA [70]	71.5±0.3
[A] CADA [97]	88.8±0.1
ARN with MDAT	91.3±0.2

TABLE 6.4: Accuracy (mean) of unsupervised domain adaptation on *Office-Home* datasets across 3 independent runs.

Method	Ar→Cl	Ar→Pr	Ar→Rw	Cl→Ar	Cl→Pr	Cl→Rw	Pr→Ar	Pr→Cl	Pr→Rw	Rw→Ar	Rw→Cl	Rw→Pr	Avg
ResNet-50 [141]	34.9	50.0	58.0	37.4	41.9	46.2	38.5	31.2	60.4	53.9	41.2	59.9	46.1
DAN [3]	43.6	57.0	67.9	45.8	56.5	60.4	44.0	43.6	67.7	63.1	51.5	74.3	56.3
DANN (DAT) [2]	45.6	59.3	70.1	47.0	58.5	60.9	46.1	43.7	68.5	63.2	51.8	76.8	57.6
TransNorm+DANN [153]	43.5	60.9	72.1	51.0	61.5	62.5	49.6	46.8	70.4	63.7	52.2	77.9	59.3
CDAN [145]	49.0	69.3	74.5	54.4	66.0	68.4	55.6	48.3	75.9	68.4	55.4	80.5	63.8
ARN with MDAT (Proposed)	51.3	69.7	76.2	59.5	68.3	70.0	57.2	48.9	75.8	69.1	55.3	80.6	65.2

FIGURE 6.3: The training procedure *w.r.t.* loss and test accuracy. $\mathcal{L}_e$ is the training loss of reconstructing target samples in Eq. 6.6. $\mathcal{L}_r$ is the training loss of reconstruction network in Eq. 6.4. $\mathcal{L}_d$ is the domain loss in DAT [2]. α is the penalty term of $\mathcal{L}_e$ and $\mathcal{L}_d$ in MDAT and DAT, respectively.

6.3.3 Analyses

Ablation study. To verify the contribution of the reconstruction network G_r and **MDAT**, we discard the term $\mathcal{L}_r(\mathbf{x}^t)$ in Eq.(6.4), and evaluate the method, denoted as **ARN w.o. MDAT** in Table 6.2. Comparing it with source-only model, we can infer the improvement of reconstructing target samples. **ARN w.o. MDAT** improves tasks with low-level *domain shift* such as **MNIST** $\leftrightarrow$ **USPS**, which conforms with our discussion that unsupervised reconstruction is instrumental in learning pixel-level features. Comparing **ARN w.o. MDAT** with the original **ARN**, we can infer the contribution of **MDAT**. Table 6.2 shows that the **MDAT** achieves an impressive margin-of-improvement. For **USPS** $\rightarrow$ **MNIST** and **SVHN** $\rightarrow$ **MNIST**, the MDAT improves **ARN w.o. MDAT** by around 30%. It demonstrates that MDAT that helps learn domain-invariant features is the main reason for the tremendous improvement.

Toy dataset. We study the behavior of MDAT on a variant of *inter-twinning moons* 2D problem, where the target samples are rotated 30° from the source samples. 300 samples are generated for each domain using **scikit-learn** [155]. The adaptation ability is investigated by comparing MDAT with DANN and source-only model. As shown in Figure 6.2, we visualize the changing boundaries during 10 epochs training. In Figure 6.2(a), the model is overfitting the source domain, and the decision boundary does not change. In Figure 6.2(b) and 6.2(c), both DANN and MDAT adapt the boundaries to the target samples, but MDAT shows

TABLE 6.5: Accuracy (%) on **SVHN**→**MNIST**.

α	0.01	0.03	0.07	0.1	0.2	0.3	0.5	1.0
DANN	71.1	74.1	72.7	74.1	74.7	9.6	9.7	10.3
ARN ($m = 1$)	95.7	95.9	93.3	93.2	80.1	75.3	73.1	67.5
m	0	0.1	0.3	0.5	1.0	2.0	5.0	10.0
ARN ($\alpha = 2e^{-2}$)	64.3	64.5	75.2	90.0	96.0	97.4	97.7	96.7

TABLE 6.6: Visualizing the source image, target images and reconstructed target images (R-Target Images) for four digit adaptation tasks.

	Source Images	Target Images	R-Target Images
MNIST → USPS
USPS → MNIST
SVHN → MNIST
SYN → SVHN

faster and better adaptation during 10 epochs. Integrating the training procedure of **SVHN**→**MNIST** in Figure 6.3(a), we justify that more effective gradients are provided by MDAT for better adaptation performance.

Gradients and stability analysis. We further study the training procedure of MDAT on **SVHN**→**MNIST** *w.r.t.* loss and target accuracy in Figure 6.3(a) and 6.3(b), respectively. In Figure 6.3(a), ARN has steadily decreasing loss ($\mathcal{L}_r$) for all α , but the domain loss in DAT ($\mathcal{L}_d$) becomes extremely small at the beginning. These observations conform with our intuition: the domain classifier in DAT is too strong to impede the adversarial training, while MDAT provides more effective gradients for training feature extractor by restricting the capacity of domain classifier. With effective gradients, the adversarial game is more balanced, which is validated in Figure 6.3(b) where the test accuracy of ARN is more stable than that of DAT across training epochs.

Parameter sensitivity. We investigate the sensitivity of the hyper-parameters α and m on **SVHN**→**MNIST**. In Table 6.5, the results show that ARN achieves

FIGURE 6.4: T-SNE visualization on **SVHN**→**MNIST** with their corresponding domain labels (red: target; blue: source) and category labels (10 classes) shown in the left and right subfigures, respectively.

good performance as $\alpha \in [0.01, 0.1]$. Even with larger α , ARN is able to achieve convergence. In comparison, denoting α as the weight of adversarial domain loss ($\mathcal{L}_d$), the DANN cannot converge when $\alpha > 0.2$ due to the imbalanced adversarial game between the overwhelming domain classifier and the feature extractor. For the sensitivity of m , the accuracy of ARN exceeds 96.0% as $m \geq 1$. In Section 6.2.5, as $m \geq 1$, the decoder serves as a domain classifier. These analyses validate that ARN is more insensitive to the parameters than that of DANN. Even in the worst cases, ARN can always achieve convergence.

T-SNE embeddings. We analyze the performance of domain alignment for DANN (DAT) [2] and ARN (MDAT) by plotting T-SNE embeddings of the features $\mathbf{z}$ on the task **SVHN**→**MNIST**. In Figure 6.4(a), the source-only model obtains diverse embeddings for each category but the domains are not aligned. In Figure 6.4(b), the DANN aligns two domains but the decision boundaries of the classifier are vague. In Figure 6.4(c), the proposed ARN effectively aligns two domains for all categories and the classifier boundaries are much clearer.

Interpreting adapted features via reconstruction. One of the key advantages of ARN is that by visualizing the reconstructed target images we can infer how the features are domain-invariant. We reconstruct the MDAT features of the target domain and visualize them in Table 6.6. It is observed that the target features are reconstructed to source-like images by the decoder G_r . As discussed before, intuitively, MDAT forces the target feature to mimic the source feature distribution, which conforms with the visualization. Similar to image-to-image translation, this indicates that our method conducts implicit feature-to-feature translation that transfers the target features to source-like features, and hence the features are domain-invariant.

6.4 Conclusion

We proposed a new domain alignment approach namely Max-margin Domain-adversarial Training (MDAT) and a MDAT-based deep neural network for unsupervised domain adaptation. The proposed method offers effective and stable gradients for feature learning via an adversarial game between the feature extractor and the reconstruction network. The theoretical analysis provides justifications on how it minimizes the distribution discrepancy. Extensive experiments demonstrated the effectiveness of our method and we further interpreted the features by visualization that conforms with our insight. Potential evaluation on semi-supervised learning constitutes our future work.

Chapter 7

Generalized MDAT: Asymmetric Adversarial Domain Adaptation

This chapter studies the transferability and discriminability of deep features in domain adaptation. From this perspective, we generalize MDAT to a more generic form: Asymmetric Adversarial Domain Adaptation (AADA). Section 7.1 introduces the motivation and the connections with MDAT. Section 7.2 revisits adversarial domain adaptation while highlighting its limitations. Then the AADA method is detailed in Section 7.3. Section 7.4 demonstrates the effectiveness and superiority of AADA. Section 7.5 concludes the chapter.

7.1 Introduction

As proposed in Chapter 6, MDAT achieves remarkable and stable results. The principles of MDAT and DAT are quite similar — the minimization of the discrepancy term in Theorem 2.3.1, but the MDAT yields better performance than DAT on all benchmarks. The better performance is resulted from the pixel-level alignment by the unsupervised reconstruction. Besides, the asymmetric form might be a crucial factor. In this chapter, we discuss this problem from the perspectives of transferability and discriminability of features across domains, and generalize MDAT to a asymmetric adversarial form.

Prevailing domain adaptation [69] tackles the problem of *domain shift* by enhancing the transferability of feature learning, *i.e.* aligning the marginal distributions across domains. For deep neural networks, aligning the deep features mainly relies on two categories of approaches. One category reduces the distribution discrepancy by measuring the statistics [58, 61], which is simple to implement and usually possesses stable convergence. Another category is adversarial domain adaptation, inspired by the Generative Adversarial Network (GAN) [63]. A binary domain discriminator is introduced to distinguish the domain labels while the feature extractor learns to fool the discriminator in DAT [2]. Adversarial domain adaptation approaches have achieved prominent performance on many challenging tasks including semantic segmentation [156], sentiment classification [157] and wireless sensing [97].

Though adversarial domain adaptation methods yield superior performance, they bring “uncertainty” to feature learning [71]. Such uncertainty is due to the side effect of domain adversarial training (DAT). Specifically, any features that can deceive the domain discriminator and perform well in the source domain conform with the goal of DAT. The worst consequence could be overfitting to the source domain and generating meaningless features in the target domain as long as they can fool the domain discriminator. Therefore, this uncertainty in adversarial training is severely detrimental to learning discriminative features in the target domain. This explains why many subsequent works aim to preserve semantic information [74] or adjust the boundary of classifier during DAT [72, 73]. These solutions ameliorate the traditional DAT yet by adding more learning steps, which either increases the computational overhead or requires a sophisticated hyper-parameter tuning process for multiple objectives.

In this chapter, we address the uncertainty problem by proposing an Asymmetric Adversarial Domain Adaptation (AADA). The key problem of DAT consists in the symmetric objective which is to equally push two domains as close as possible, deteriorating feature learning and neglecting the decision boundary. Inspired by the fact that the source domain has great discriminability due to full supervision, AADA aims to fix the source domain and only adapts the target domain. To this end, we design an autoencoder serving as a domain discriminator to embed the source features, while a feature extractor is trained to deceive it to embed the target features. Such adversarial process can be realized by a minimax game with

FIGURE 7.1: Comparison between previous Domain Adversarial Training (DAT) method and ours. **Left:** The discrepancy region is large before adaptation. **Middle:** DAT based on a domain classifier aligns two domains together and pushes them as close as possible, which hurts the feature discriminability in the target domain. **Right:** AADA fixes the source domain, regarded as the low-energy space, and pushes the target domain to approach the source one, which makes use of the well-trained classifier in the labeled source domain.

a margin loss. As shown in Figure 7.1, DAT employs a binary domain discriminator to align two domains together, while we only force the target domain to approach the source domain that possesses good discriminability. Furthermore, it is acknowledged that the autoencoder is an energy function that learns to map the observed sample to the low-energy space [158]. Therefore, energy function can cluster similar data to form a high density manifold, which helps preserve more semantic information. We leverage the energy function to fix the source domain by associating lower energies to it while pushing the target domain to the low-energy space via adversarial training in an innovative manner.

This also explains why MDAT performs better in Chapter 6. As the reconstruction network serves as the domain discriminator, the adversarial training is also asymmetric, which helps preserve the discriminability of features. AADA further generalizes the domain discriminator of MDAT to be an energy function, and as long as the asymmetric training objectives are employed in DAT, better discriminability can be achieved.

The proposed AADA is a novel and new fundamental domain alignment technique which can be integrated with other domain adaptation approaches easily. Our contribution in this chapter is threefold:

- We propose a novel asymmetric adversarial scheme that replaces the conventional domain classifier with an autoencoder, which incorporates only the target domain into the adversarial feature training, circumventing the loss of discriminability in traditional domain adversarial training.
- The autoencoder is an energy function that maps the two domains to the low-energy space, which encourages the feature clusters to be tight and thus further benefits the classification task in an unsupervised manner. Theoretically, any energy functions work with AADA.
- AADA is a generic domain alignment approach that can be used as an ingredient in existing domain adaptation approaches. The experiment validates that AADA outperforms other domain alignment approaches significantly. We further demonstrate the boosted discriminability by spectral analysis.

7.2 Revisiting Transferability and Discriminability in Symmetric Adversarial Domain Adaptation

We first revisit the learning theory of unsupervised domain adaptation (UDA) [48] where we analyze how the target expected error $\epsilon_T(h)$ is connected to transferability and discriminability in representation learning. Then we highlight the problem of forfeiting discriminability due to symmetric adversarial training, and excogitate the insights of our approach that conceals behind.

7.2.1 The Theory of Domain Adaptation

In unsupervised domain adaptation, we have access to N_s labeled examples from a source domain $\mathcal{D}_S = \{\mathbf{x}_i^s, y_i^s | \mathbf{x}_i^s \in \mathbf{X}_s, y_i^s \in Y_s\}$ and N_t unlabeled examples from a target domain $\mathcal{D}_T = \{x_i^t | \mathbf{x}_i^t \in \mathbf{X}_t\}$, which are sampled from distinct distributions $\mathbb{P}$ and $\mathbb{Q}$, respectively. The objective of UDA is to learn a model that performs well for the target domain. The learning theory of UDA was proposed by Ben-David [48] that has been reviewed in Theorem 2.3.1. Let $\mathcal{H}$ be the common hypothesis class for source and target. The expected error for the target domain ϵ_t is upper

bounded by the source error ϵ_s , the divergence term $\frac{1}{2}d_{\mathcal{H}\Delta\mathcal{H}}(\mathcal{D}_S, \mathcal{D}_T)$, the ideal risk λ and a constant of model complexity C_0 .

As source data is annotated, the source error $\epsilon_s(h)$ can be simply minimized via supervised learning. To minimize $\epsilon_t(h)$, UDA focuses on reducing the domain discrepancy term $\frac{1}{2}d_{\mathcal{H}\Delta\mathcal{H}}(\mathcal{D}_S, \mathcal{D}_T)$ and the ideal risk λ . In representation learning for UDA, minimizing the domain discrepancy is able to improve **transferability** of features. This can be achieved by domain adversarial training [2] or the minimization of statistical measures of such discrepancy [58]. Another criterion that plays a vital role in feature representations is **discriminability**. It refers to the capacity of clustering in the feature manifold, and therefore controls the easiness of separating categories. For UDA, we pursue good discriminability in both source and target domains simultaneously. As we typically use a shared feature extractor for two domains, enhancing discriminability is equivalent to seeking for a better ideal joint hypothesis $h^* = \min_h[\epsilon_s(h) + \epsilon_t(h)]$ [71].

7.2.2 Limitations and Insights

From the analyses above, the feature learning of UDA should guarantee both transferability and discriminability, which inspires us to investigate the existing domain alignment methods from these two perspectives. Adversarial domain adaptation methods have shown prominent performance and become increasingly popular. Ganin *et al.* [2] pioneered it by proposing Domain Adversarial Neural Network (**DANN**) that learns domain-invariant features using the reverse gradients from a domain classifier. Typically, adversarial UDA approaches consist of a shared feature extractor $\mathbf{f} = G_f(x)$, a label predictor $\mathbf{y} = G_y(x)$ and a domain discriminator $\mathbf{d} = G_d(x)$. Apart from standard supervised learning on the source domain, a minimax game between $\mathbf{f}$ and $\mathbf{d}$ is designed. The domain discriminator $\mathbf{d}$ is trained to distinguish the domain label between the source domain and the target domain, while the feature extractor $\mathbf{f}$ learns to deceive the domain classifier $\mathbf{d}$. In this manner, the domain adversarial training enables the model to learn transferable features across domains when the Nash Equilibrium is achieved. The whole process

can be formulated as

$$\min_{G_f, G_y} \mathcal{L}_c(\mathbf{X}_s, Y_s) - \gamma \mathcal{L}_c(\mathbf{X}_s, \mathbf{X}_t), \quad (7.1)$$

$$\min_{G_d} \mathcal{L}_c(\mathbf{X}_s, \mathbf{X}_t), \quad (7.2)$$

where $\mathcal{L}_c$ is the classification loss such as cross-entropy loss, and the hyper-parameter γ decides the importance of transferability in feature learning.

Symmetric DAT does improve the transferability across domains but sacrifices the discriminability in the target domain. Let us analyze two objectives of training feature extractor $\mathbf{f}$: (1) good discriminability in the source domain and (2) learning representations that are indistinguishable to the domain discriminator. There is no constraint on the discriminability in the target domain. As depicted in Figure 7.1, DAT is symmetric and makes two domains as close as possible. Theoretically, the worst case is that the feature extractor generates meaningless representations on the target domain, as long as they can deceive the domain classifier. Hence, a good decision boundary on the source domain cannot perform well on the target domain. Previous works quantified the discriminability by spectral analysis and drew a similar conclusion [71, 149].

We believe that *the decreasing discriminability is caused by symmetric adversarial training that involves both source domain and target domain in adversarial feature learning* as shown in the second term of Eq.(7.1). Specifically, to deceive the binary domain discriminator, DAT aims to push two domains close. In this process, DAT cannot control how the domains are aligned and cannot guarantee whether the decision boundary separates the categorical clusters in the target domain. As depicted in Figure 7.1, this problem can be overcome by the proposed asymmetric adversarial training. In our approach, we fix the source domain and only render the target domain to approach the source. In this fashion, the feature discriminability is preserved and a good classifier is easily obtained. To achieve symmetric adversarial training, we innovatively propose to leverage the autoencoder as the domain classifier.

FIGURE 7.2: AADA model constitutes a shared feature extractor G_f , a classifier G_y and an autoencoder G_a . Except for supervised learning on the source domain, the autoencoder plays a domain discriminator role that learns to embed the source features and push the target features away, while the feature extractor learns to generate target features that can deceive the autoencoder. Such process is an asymmetric adversarial game that pushes the target domain to the source domain in the feature space.

7.3 Asymmetric Adversarial Domain Adaptation

Maintaining the source manifolds during DAT is not a trivial task with the consideration of the complexity of network architecture. This requires us to design a simple yet effective asymmetric adversarial mechanism. In this section, Asymmetric Adversarial Domain Adaptation (AADA) is detailed and we theoretically justify how it reduces domain discrepancy.

7.3.1 The Learning Framework

As shown in Figure 7.2, our model consists of a shared feature extractor G_f parameterized by θ_f , a label predictor G_y parameterized by θ_y and an autoencoder G_a parameterized by θ_a . The feature extractor G_f , typically composed of multiple convolutional layers, embeds an input sample to a feature embedding z , and then the label predictor G_y that usually consists of several fully connected layers maps the feature embedding to the predicted label $\hat{y}$. The autoencoder G_a reconstructs an embedding z to $\hat{z}$.

In the learning phase, the first objective of the model is to learn feature discriminability in the source domain. As we have access to the labeled source data $(\mathbf{X}_s, Y_s)$, it is simply achieved by minimizing the cross-entropy loss via back-propagation:

$$\min_{G_f, G_y} \mathcal{L}_{CE}(\mathbf{X}_s, Y_s) = -\mathbb{E}_{(\mathbf{x}_s, y_s) \sim (\mathbf{X}_s, Y_s)} \sum_{n=1}^{N_s} [\mathbb{I}_{[l=y_s]} \log G_y(G_f(\mathbf{x}_s))]. \quad (7.3)$$

With robust feature learning in the source domain, the next objective is to learn transferable representations using the unlabeled data $(\mathbf{X}_t)$ in the target domain. To this end, we propose an asymmetric adversarial training scheme that involves an autoencoder G_a with a margin Mean Squared Error (MSE) loss. The autoencoder G_a , which plays a domain discriminator role, only learns to embed features from the source domain, but not to embed features from the target domain. The objective of the autoencoder-based domain discriminator is formulated as:

$$\min_{G_a} \mathcal{L}_{AE}(\mathbf{X}_s) + \max(0, m - \mathcal{L}_{AE}(\mathbf{X}_t)), \quad (7.4)$$

where m is the margin between two domains in the feature space. Here, the MSE loss of the autoencoder is defined as:

$$\mathcal{L}_{AE}(\mathbf{x}_i) = \|G_a(G_f(\mathbf{x}; \theta_f); \theta_a) - \mathbf{x}_i\|_2^2, \quad (7.5)$$

where $\|\cdot\|_2^2$ denotes the squared L_2 -norm.

To play the adversarial game, the feature extractor G_f learns to fool the autoencoder G_a by generating source-like features for the samples in the target domain. When the feature extractor succeeds, the representations of the target domain can inherit good discriminability from the source domain, and the label predictor G_y trained in the source domain applies equally. The adversarial training of the feature extractor G_f is formulated as:

$$\min_{G_f} \mathcal{L}_{AE}(\mathbf{X}_t). \quad (7.6)$$

The overall optimization of the proposed AADA model is formally defined by

$$\begin{aligned} & \min_{G_f, G_y} \mathcal{L}_{CE}(\mathbf{X}_s, Y_s) + \gamma \mathcal{L}_{AE}(\mathbf{X}_t), \\ & \min_{G_a} \mathcal{L}_{AE}(\mathbf{X}_s) + \max(0, m - \mathcal{L}_{AE}(\mathbf{X}_t)), \end{aligned} \quad (7.7)$$

where γ is a hyper-parameter that controls the importance of transferability. In this fashion, our approach only incorporates the $\mathcal{L}_{AE}(\mathbf{X}_t)$ term into the training of G_f , which pushes the target domain to the source domain. Oppositely, the objective of G_a serves as a domain discriminator that pushes two domains away by a margin m .

7.3.2 Discussions and Theories

The autoencoder used in our approach is an energy function, which associates lower energies (*i.e.* MSE) to the observed samples in a binary classification problem [158]. For UDA, the autoencoder in our model associates low energies to the source features, and AADA compels the target features to approach the low-energy space. The design of such adversarial scheme is inspired by the Energy-based GAN which theoretically proves that using an energy function in GAN, the true distribution can be simulated by the generator at *Nash Equilibrium* [152]. Similarly, in AADA, the feature extractor G_f can mimic the source distribution for the samples in the target domain when the model achieves convergence.

From the formulation, AADA generalizes MDAT with respect to the training procedure and network architecture. Firstly, the reason why MDAT outperforms conventional DAT is that the asymmetric adversarial training preserves discriminability. It is further demonstrated in AADA that preserving the source manifold is important for good discriminability. Secondly, such asymmetric form is achieved by the reconstruction network in MDAT and auto-encoder in AADA, both of which significantly improve DAT. Theoretically, the autoencoder can be replaced by many energy functions. Hence, AADA might be further improved by replacing the autoencoder with other energy functions with asymmetric objectives.

How AADA minimizes the discrepancy is the same as MDAT. As theoretically justified in [2], domain adversarial training using a domain classifier effectively reduces the domain discrepancy term $d_{\mathcal{H}\Delta\mathcal{H}}(\mathcal{D}_S, \mathcal{D}_T)$ in Theorem 2.3.1. In AADA,

the autoencoder can be treated as a form of a domain classifier with a margin. The MSE loss of the autoencoder in our framework, *i.e.* $\mathcal{L}_{AE}(\cdot)$, works in the same way as the domain classifier. The training objective of $\mathcal{L}_{AE}(\cdot)$ is $\mathcal{L}_{AE}(\mathbf{X}_s) = 0$ and $\mathcal{L}_{AE}(\mathbf{X}_t) = m$, which is equivalent to the functionality of the standard domain classifier. Therefore, following the theorem proved in [2], the proposed autoencoder can maximize the domain divergence while the adversarial feature learning minimizes the divergence by deceiving the autoencoder. Moreover, as the autoencoder memorizes more domain information than a binary classifier, it transfers more knowledge during asymmetric adversarial training.

7.4 Experiments

We evaluate the proposed AADA on three UDA benchmarks and compare our model with the prevailing approaches. Then we validate the improved feature discriminability by spectral analysis. In the discussions, we verify our motivations by quantitatively analyzing the transferability and discriminability.

7.4.1 Experimental Setup

We evaluate all five transfer tasks on **Digits** [93] with multiple levels of domain difference. We follow the experimental settings of **DANN** [2]. Furthermore, AADA is evaluated on **Image-CLEF**, **Office-Home** and **VisDA-2017**, which have been illustrated in Chapter 2. The protocol for these three datasets is the full-training protocol in **DANN** [2].

Baselines. We compare our AADA with the state-of-the-art UDA methods: **MMD** [58], Deep Adaptation Network (**DAN**) [3], Deep **CORAL** [61], **DANN** [2], **JAN** [140], Deep Reconstruction Classification Network (**DRCN**) [89], Domain Separate Network (**DSN**) [90], **ADDA** [70], **CoGAN** [159], **CyCADA** [66], Maximum Classifier Discrepancy (**MCD**) [73], Conditional Domain Adversarial Network (**CDAN**) [145], Batch Spectral Penalization (**BSP**) [71], **CCN** [80], **GTA** [160] and **MCS** [161]. Among these methods, **MMD**, **DAN**, **JAN**, **D-CORAL**, **DANN** and our AADA are only based on domain alignment. To further

prove that our method can be integrated with other novel methods to achieve better adaptation, we integrate our method with **CCN** [80] that employs self-training to improve transferability of feature learning. We compare **AADA+CCN** with the recent state-of-the-art methods, including **MCD** that employs the classifier discrepancy, **CyCADA** that introduces the cyclic loss of two domains, **GTA** that incorporates a sample generator, and **BSP** that emphasizes the importance of discriminability in domain adaptation.

Implementation Details. For the task with 32×32 images, we use the same network architecture as DANN, while for **Image-CLEF** and **Office-Home**, we use **ResNet-50** with pretrained parameters on ImageNet [46]. In AADA, we use an autoencoder with only fully connected layers, and the detailed network architectures are shown in Table 7.1 and 7.2. Following former UDA works [70], we use Adam optimizer with the constant learning rate $\mu = 5e^{-4}$ for digit adaptation, and SGD with the decaying learning rate for object recognition. For hyper-parameters m and γ , we set $m = 0.5, \gamma = 1e^{-2}$ for 32×32 images, and $m = 1, \gamma = 1e^{-1}$ for object recognition datasets, which are empirically obtained by cross-validation on MNIST→USPS and Image-CLEF. The whole experiment is implemented by **PyTorch** framework.

TABLE 7.1: The network architecture used in digit adaptation .

Layer Index	Feature Extractor	Autoencoder	Label Predictor
0	$32 \times 32 \times 3$ Image	-	-
1	5×5 conv. 64 ReLU	2048 dense, ReLU	10 dense, softmax
2	3×3 max-pool, stride 2	512 dense, ReLU	
3	5×5 conv. 64 ReLU	2048 dense, ReLU	
4	3×3 max-pool, stride 2		
5	5×5 conv. 128 ReLU		
6	3072 dense, dropout, ReLU		
7	2048 dense, dropout ReLU		

7.4.2 Overall Results

We first compare our methods with MMD, CORAL, DANN, DAN and JAN that are only based on domain alignment. In Table 7.3, our approach shows significant

TABLE 7.2: The network architecture used in Office-Home and VisDA-17.

Layer Index	Feature Extractor	Autoencoder	Label Predictor
0	$224 \times 224 \times 3$ Image	-	-
1	ResNet50 CNNs	2048 dense, ReLU	10 dense, softmax
2		512 dense, ReLU	
3		128 dense, ReLU	
4		512 dense, ReLU	
5		2048 dense, ReLU	

TABLE 7.3: Accuracy (mean $\pm$ std) on five *Digit* benchmarks for domain adaptation across 3 independent runs.

Source Target	MNIST USPS	USPS MNIST	SVHN MNIST	SYN-DIGIT SVHN	MNIST MNIST-M
Source-only	78.2	63.4	54.9	86.7	56.3
<i>Domain Alignment Methods</i>					
MMD [58]	81.1	-	71.1	88.0	76.9
CORAL [61]	80.7	-	63.1	85.2	57.7
<i>Adversarial Training based Methods</i>					
DANN [2]	85.1	73.0	74.7	90.3	76.8
CoGAN [159]	91.2	89.1	-	-	62.0
ADDA [70]	89.4	90.1	76.0	-	-
CDAN [145]	93.9	96.9	88.5	-	-
GTA [160]	95.3	90.8	92.4	-	-
CyCADA [66]	95.6	96.5	90.4	-	-
<i>Other State-of-the-Art Methods</i>					
DRCN [89]	91.8	73.7	82.0	87.5	68.3
DSN [90]	91.3	-	82.7	91.2	83.2
MCD [73]	96.5	94.1	96.2	-	-
BSP+DANN [71]	94.5	97.7	89.4	-	-
BSP+CDAN [71]	95.0	98.1	92.1	-	-
MCS+GTA [161]	97.8	98.2	91.7	-	-
AADA _{opt}	95.6 $\pm$ 0.3	92.7 $\pm$ 0.5	74.8 $\pm$ 0.8	88.8 $\pm$ 0.2	47.1 $\pm$ 1.2
AADA (Ours)	98.4$\pm$0.3	98.6$\pm$0.3	98.1$\pm$0.5	92.2$\pm$0.4	95.5$\pm$0.2

improvement against the standard DAT (DANN), even outperforming the state-of-the-art methods that require much higher computation overhead such as CyCADA and GTA due to the training of cyclic or generative networks. For pixel-level domain shift in **MNIST** $\rightarrow$ **MNIST-M**, the traditional domain alignment methods such as MMD and DANN cannot effectively deal with them, but AADA achieves

TABLE 7.4: Accuracy (mean) of unsupervised domain adaptation on *Image-CLEF* datasets across 3 independent runs.

Method	I→P	P→I	I→C	C→I	C→P	P→C	Avg
ResNet-50 [141]	74.8	83.9	91.5	78.0	65.5	91.2	80.7
DAN [3]	74.5	82.2	92.8	86.3	69.2	89.8	82.5
DANN [2]	75.0	86.0	96.2	87.0	74.3	91.5	85.0
CCN [80]	77.1	87.5	94.0	86.0	74.5	91.7	85.1
JAN [140]	76.8	88.0	94.7	89.5	74.2	91.7	85.8
CDAN [145]	76.7	90.6	97.0	90.5	74.5	93.5	87.1
AADA	78.0	90.3	94.0	87.8	75.2	93.5	86.5
AADA+CCN	79.2	92.5	96.2	91.4	76.1	94.7	88.4

TABLE 7.5: Accuracy (mean) of unsupervised domain adaptation on *Office-Home* datasets across 3 independent runs.

Method	Ar→Cl	Ar→Pr	Ar→Rw	Cl→Ar	Cl→Pr	Cl→Rw	Pr→Ar	Pr→Cl	Pr→Rw	Rw→Ar	Rw→Cl	Rw→Pr	Avg
ResNet-50 [141]	34.9	50.0	58.0	37.4	41.9	46.2	38.5	31.2	60.4	53.9	41.2	59.9	46.1
DAN [3]	43.6	57.0	67.9	45.8	56.5	60.4	44.0	43.6	67.7	63.1	51.5	74.3	56.3
DANN [2]	45.6	59.3	70.1	47.0	58.5	60.9	46.1	43.7	68.5	63.2	51.8	76.8	57.6
JAN [140]	45.9	61.2	68.9	50.4	59.7	61.0	45.8	43.4	70.3	63.9	52.3	76.8	58.3
CCN [80]	47.3	65.2	70.1	51.3	60.5	60.9	48.1	45.5	71.3	65.1	53.5	77.0	59.7
CDAN [145]	49.0	69.3	74.5	54.4	66.0	68.4	55.6	48.3	75.9	68.4	55.4	80.5	63.8
BSP+DANN [71]	51.4	68.3	75.9	56.0	67.8	68.8	57.0	49.6	75.8	70.4	57.1	80.6	64.9
AADA	52.3	69.5	76.3	59.7	68.2	70.2	58.2	48.9	75.9	69.1	54.3	80.5	65.3
AADA+CCN	54.0	71.3	77.5	60.8	70.8	71.2	59.1	51.8	76.9	71.0	57.4	81.8	67.0

95.5% accuracy. Moreover, for SVHN→MNIST with larger domain shift, AADA surpasses DANN by 23.4% and BSP+DANN by 8.7%. For more challenging tasks in Table 7.4 and 7.5, the proposed AADA surpasses all the methods that are purely based on domain alignment. AADA outperforms DANN by 1.5% on Image-CLEF and 7.7% on OfficeHome. Since the domain shift is small in Image-CLEF, the improvement margin is not large.

AADA is a generic domain alignment approach that can be integrated to other novel UDA frameworks, achieving better performance. We integrate AADA with CCN [80] and evaluate it on challenging tasks. As shown in Table 7.4 and 7.5, the proposed AADA+CCN achieves the state-of-the-art accuracies, outperforming CDAN and BSP that also aim to improve discriminability. We can see that AADA improves CCN by 2.9% on Image-CLEF and 7.3% on Office-Home, which implies that AADA can bring much improvement to existing UDA methods by preserving more discriminability in domain alignment. Table 7.6 presents the accuracies of each category in the largest UDA benchmark **VisDA-2017**. On average, AADA improves MMD and DANN by 9.1% and 6.0%, respectively.

TABLE 7.6: Accuracy (%) on VisDA-2017 for domain adaptation.

Method	Accuracy
ResNet-50 [141]	53.2
MMD [58]	55.5
DANN [2]	58.6
DAN [3]	59.8
AADA	64.6

7.4.3 Spectral Analysis using SVD

The intuition and the theoretical analysis validate that the proposed model achieves a good trade-off between transferability and discriminability. Here we employ the quantitative method to further demonstrate it. The prior research proposed to apply Singular Value Decomposition (SVD) to the representation z , and then infer transferability and discriminability by the singular values and the corresponding angles of eigenvectors, respectively [71]. Motivated by this, we conducted an experiment on a digit adaptation task **SVHN**→**MNIST**. Using the feature extractor G_f , we obtain the target feature matrix $\mathbf{F}_t = [\mathbf{f}_t^1 \dots \mathbf{f}_t^b]$ where b is the batch size. Then we apply SVD to the target feature matrix as follows:

$$\mathbf{F}_t = \mathbf{U}_t \mathbf{\Sigma}_t \mathbf{V}_t^T, \quad (7.8)$$

where $\mathbf{\Sigma}_t$ denotes the eigenvalue, $\mathbf{U}_t$ denotes the set of eigenvectors and $\mathbf{V}_t$ is an unitary matrix. SVD is also applied to the source feature matrix to obtain $\mathbf{\Sigma}_s$ and $\mathbf{U}_s$. In Figure 7.3(a), we plot the top-20 normalized singular values $\mathbf{\Sigma}_t$ *w.r.t* three models including the source-only model, DANN and our AADA model. It is observed that the largest singular value of the DANN target feature matrix is greater than the other values, which impairs the semantic information included in other smaller singular values. In comparison, the distribution of singular values for AADA is similar to that for the source-only model that preserves more discriminability in feature learning. In Figure 7.3(b), we show the normalized corresponding angles of singular values. The corresponding angle depicts the commonality between the source eigenvectors $\mathbf{U}_s$ and the target eigenvectors $\mathbf{U}_t$, which indicates the transferability of the features. For DANN, the sharp distribution of the angles indicates that DANN only utilizes several peak transferable features. This deteriorates the informative representation in the target domain. Whereas,

FIGURE 7.3: Measures of discriminability and transferability.

AADA obtains more transferable feature which also shows better discriminability and has good repercussion for learning a common decision boundary.

7.4.4 Analytics and Discussions

Opposite Direction of AADA. The proposed AADA fixes the source domain and then learns to force the target domain to approach to it, which utilizes the good classifier of the source domain. What if we fix the target domain and force the source domain to approach it? We denote this situation as the opposite form of AADA, namely $AADA_{opt}$. The optimization procedure of $AADA_{opt}$ is written

as:

$$\begin{aligned} & \min_{G_f, G_y} \mathcal{L}_{CE}(\mathbf{X}_s, Y_s) + \gamma \mathcal{L}_{AE}(\mathbf{X}_s), \\ & \min_{G_a} \mathcal{L}_{AE}(\mathbf{X}_t) + \max(0, m - \mathcal{L}_{AE}(\mathbf{X}_s)). \end{aligned} \quad (7.9)$$

Intuitively, as the decision boundary of the target domain keeps changing during training, this may weaken the discriminability in domain adversarial training, which is similar to DANN. The result in Table 7.3 verifies our analysis. AADA_{opt} only produces similar results as DANN, which implies that it is infeasible to align two domains and the boundaries of classifier simultaneously.

FIGURE 7.4: The training stability and the sensitivity of the hyper-parameters.

FIGURE 7.5: Accuracy by varying m (left) and γ (right).

Sensitivity Study. To demonstrate that AADA is not sensitive to the hyper-parameters, we conduct the experiments on **SVHN**→**MNIST** across multiple m and γ . Figure 7.5(a) illustrates the sensitivity results in terms of the margin m . As the margin increases, the accuracy increases until 0.5, which conforms with our intuition. When the margin is small, the degree of transferability is limited. If the margin is greater than 0.1, we observe that the results are very stable in Figure 7.4(b). As to γ , as illustrated in Figure 7.5(b), we compare our approach with Domain Adversarial Training (DAT) which is the reproduced DANN where γ is the

weight of the adversarial loss in Eq.(7.1). In Figure 7.5(b) and Figure 7.4(c), the results of AADA are robust given large γ but DANN cannot converge with large γ . In Figure 7.4(a), we can observe that the target accuracies of DAT fluctuate a lot during training while AADA provides a more stable training procedure. The sensitivity study is consistent with the insights provided earlier, and moreover the results show more robustness *w.r.t* training procedure, which makes it easier to apply our method in an unsupervised manner.

Ideal Joint Hypothesis. We estimate the ideal joint hypothesis λ in Theorem 2.3.1 to show the **discriminability** of feature embeddings. To this end, we train an MLP classifier on all source and target data with labels. As shown in Figure 7.3(c), AADA has the lowest λ in the feature space on task **SVHN**→**MNIST**. This demonstrates that AADA has satisfactory discriminability in the feature space of two domains, which helps learn a good decision boundary that separates the data from two domains.

Distribution Discrepancy. As proposed in the theory of domain adaptation [48], the A -distance is a measure of domain discrepancy that quantifies the **transferability** of feature embeddings. It is defined as $d_A = 2(1 - 2\epsilon)$ where ϵ is the error of a domain classifier. We train an MLP classifier to discriminate source and target domains on task **SVHN**→**MNIST**. Results are shown in Figure 7.3(d), and it is observed that the A -distance of AADA is the lowest, which indicates better transferability of AADA compared with DANN.

Visualization Representation. We visualize learned features on the *Digit* task **SVHN**→**MNIST** via t-SNE [147] and present it in Figure 7.6. The visualization validates the insights of AADA. DANN aligns the features of two domains, but we can see that due to the lack of discriminability, some categories of data are confused. Hence many approaches were proposed to adjust the boundary for DANN features [73]. In comparison, the AADA features are domain-invariant and preserve the good decision boundary of the source domain simultaneously. This further proves that AADA learns better discriminability in domain alignment.

FIGURE 7.6: The t-SNE visualization of the embeddings z on task **SVHN** $\rightarrow$ **MNIST**. The top figures are visualization with domain labels (blue: source, red: target). The bottom figures are visualization with category labels (10 classes).

7.5 Conclusion

In this chapter, we proposed a novel asymmetric adversarial regime for unsupervised domain adaptation. As the conventional adversarial UDA methods affect the discriminability while improving the transferability of feature representations, AADA aims to preserve the discriminability by encouraging the target domain to approach the source domain in the feature space, which is achieved by an autoencoder with an asymmetric adversarial training scheme. Spectral analysis was utilized to justify the improved discriminability and transferability. The experimental results demonstrated its robustness and superiority on numerous large-scale datasets.

This chapter explained why MDAT is better than DAT from the perspectives of feature learning. Specifically, MDAT is a special form of AADA where the deep reconstruction network works as the domain discriminator. The asymmetric form of

adversarial training helps MDAT preserve more discriminability than conventional DAT, which leads to better results.

Chapter 8

Imbalanced Domain Adaptation for Long-tailed Data

In this chapter, a new domain adaptation scenario — Imbalanced Domain Adaptation (IDA) is introduced. IDA considers both covariate shift and label shift simultaneously, which is commonly confronted in real-world scenarios. Section 8.1 offers real-world cases and the motivation. Then in Section 8.2, the Cluster-level Discrepancy Minimization (CDM) is proposed to tackle the IDA problem. In Section 8.3, extensive experiments are conducted using the IDA evaluation protocol which demonstrate that CDM achieves state-of-the-art performance on all settings and benchmarks. Section 8.4 concludes the chapter.

8.1 Introduction

In the last two chapters, we have studied how to stabilize the domain adaptation training and how to improve the performance, respectively. In this chapter, we aim to deal with the label shift in domain adaptation, which is commonly seen in real-world scenarios.

Though current UDA methods achieve remarkable performance by reducing the domain shift between source distribution $\mathbb{P}$ and target distribution $\mathbb{Q}$, one fundamental assumption is that the conditional output distribution is invariant, *i.e.* $p(y|x) = q(y|x)$, where x, y are the input data and output labels, respectively [12].

FIGURE 8.1: (a) Standard domain adaptation assumes that the source domain and the target domain share the same label distributions, which is followed by prevailing UDA benchmarks with balanced category. (b) Imbalanced domain adaptation considers the real-world situation when label shift exists. For example, if the source domain is an urban area there are more cars and buses, while if the target domain is a rural area there are more bikes and horses.

In fact, since the output (label) distributions $p(y)$ and $q(y)$ are usually different, this assumption may not hold for many real-world situations. One of the most common examples is that the long-tailed data summarizes the frequency distribution of the realistic visual categories [162], while most public datasets are rather balanced. Another example is that the medical diagnostic models trained on data from one area may give wrong predictions for another area with unobserved population and time frames due to such label shift (*e.g.* different disease incidence). Theoretically, covariate shift corresponds to causal learning while label shift is related to anti-causal learning [163], which indicates that they have different repercussions for practical cross-domain problems.

Towards the aforementioned situations, this chapter studies the *Imbalanced Domain Adaptation* (IDA) problem which assumes that label shift also exists in standard domain adaptation protocol, as described in Fig. 8.1. Since label shift breaks the assumption of standard domain adaptation, IDA becomes more challenging. Specifically, aligning the marginal distributions, which is usually employed in UDA, cannot be adopted for label shift due to negative transfer [164]. It is the worst case when an adapted model performs even worse than a source-only model [165]. Furthermore, one of the main reasons is that the long-tailed data deteriorates the representation learning due to the imbalanced label distribution and small sample number of certain categories [81].

In this chapter, we propose a novel domain adaptation model, namely *Cluster-level Discrepancy Minimization* (CDM), dedicated to addressing the IDA problem. The framework of CDM constitutes three steps. (1) Considering label shift, we firstly

FIGURE 8.2: The illustration of existing UDA solutions and our proposed CDM for IDA problem. Existing approaches cannot deal with label shift by direct minimization. Our approach learns discriminative clusters and then minimizes the cluster-level discrepancy with boundary refinement, overcoming the negative effect of label shift *w.r.t.* the discrepancy minimization.

conduct a cross-domain similarity learning to obtain tight and discriminative clusters for each category in the two domains. The cluster-based feature learning renders the samples of the same category to be closer in the feature space, which mitigates the effect of label shift. (2) We align the cross-domain clusters by minimizing both feature-level and distribution-level discrepancy. (3) We employ the conditional entropy maximization to encourage the classifier to keep away from data-dense region (*i.e.* clusters) in the feature space. In this manner, CDM aligns the feature distributions of the source and target data with both covariate shift and label shift, which is the key improvement over conventional UDA methods. Extensive experimental results demonstrate that CDM outperforms the state-of-the-art UDA methods on public datasets with different levels of label shift.

8.2 Imbalanced Domain Adaptation

This chapter introduces an *Imbalanced Domain Adaptation* (IDA), a real-world domain adaptation scenario where there exist covariate shift and label shift between

the source domain and the target domain. In IDA, given N_s labeled samples $\{(\mathbf{x}_i^s, y_i^s)_{i=1}^{N_s}\}$ from the source domain $\mathcal{D}_s = \{\mathcal{X}_s, \mathcal{Y}_s\}$, where $\mathbf{x}_i^s \in \mathcal{X}_s$ and $y_i^s \in \mathcal{Y}_s$, and N_t unlabeled samples $\{(\mathbf{x}_i^t)_{i=1}^{N_t}\}$ from the target domain $\mathcal{D}_t = \{\mathcal{X}_t\}$, where $\mathbf{x}_i^t \in \mathcal{X}_t$. The two domains share the same label space with K categories. There exists both covariate shift and label shift between $\mathcal{D}_s$ and $\mathcal{D}_t$ such that $p(x) \neq q(x)$ and $p(y) \neq q(y)$. Zhao *et al.* [164] proposed a lower bound of the joint risk of $\mathcal{D}_s$ and $\mathcal{D}_t$ considering label shift.

Theorem 8.2.1. [164] Denote d_{JS} as the JS-divergence. Suppose that a Markov chain $\mathcal{X} \xrightarrow{g} \mathcal{Z} \xrightarrow{f} \hat{\mathcal{Y}}$ holds and $d_{JS}(\mathcal{Y}_s, \mathcal{Y}_t) \geq d_{JS}(\mathcal{Z}_s, \mathcal{Z}_t)$, then:

$$\epsilon_s(f \circ g) + \epsilon_t(f \circ g) \geq \frac{1}{2}(d_{JS}(\mathcal{Y}_s, \mathcal{Y}_t) - d_{JS}(\mathcal{Z}_s, \mathcal{Z}_t))^2.$$

Theorem 8.2.1 implies that if marginal label shift $d_{JS}(\mathcal{Y}_s, \mathcal{Y}_t)$ is large, then minimizing the discrepancy of the marginal distribution $d_{JS}(\mathcal{Z}_s, \mathcal{Z}_t)$ will increase the target error ϵ_t . It explains why standard UDA approach may lead to negative transfer for IDA, and indicates that aligning label shift $d_{JS}(\mathcal{Y}_s, \mathcal{Y}_t)$ is the premise of feature distribution alignment $d_{JS}(\mathcal{Z}_s, \mathcal{Z}_t)$.

The objective of this chapter is to develop a deep network $h_\theta = f \circ g$ parameterized by θ , which helps to close the domain gap by learning transferable features $g(\mathbf{x})$ and adapted classifier $y = f(g(\mathbf{x}))$. Theoretically, it helps find a smaller target risk $\epsilon_t = Pr_{(\mathbf{x}, y) \sim q}[f(g(\mathbf{x})) \neq y]$ [164].

8.2.1 Methodology Overview

As illustrated in Fig. 8.2, the proposed *Cluster-level Discrepancy Minimization (CDM)* has three objectives. Firstly, a similarity-based cluster learning scheme is proposed to alleviate the effect of label shift in feature learning. Secondly, with the rough clusters, we align the cross-domain clusters and prototype-level MMD, which reduces the cluster-level discrepancy in an iterative manner. Thirdly, with tight clusters, the boundary is further refined to satisfy the cluster assumption [166].

Starting with the supervised learning in the source domain, we minimize the cross-entropy loss formulated as:

$$\mathcal{L}_y(\mathcal{X}_s, \mathcal{Y}_s; \theta) = -\mathbb{E}_{(\mathbf{x}, y) \sim \mathbb{P}}[y^T \ln h_\theta(\mathbf{x})]. \quad (8.1)$$

The supervised classification loss is to learn the mapping $\mathcal{X}_s \rightarrow \mathcal{Y}_s$ without considering domain shifts.

8.2.2 Discriminative Cluster Learning

It is known that training a model using imbalanced data leads to the situation where the categories with large sample number overwhelm the feature space and are much easier to be predicted. In IDA, imbalanced label space is inevitable such as in the case of long-tailed data. To handle this problem, we tend to tighten the feature embedding for each category. One solution is the prototype-based method that minimizes the distance between the learnable prototypes and the samples in the same category [44]. As it introduces more parameters, we propose a more efficient cluster learning scheme that is inspired by similarity learning.

For each pair of samples in a mini-batch, an indicator $\kappa(\mathbf{x}_i, \mathbf{x}_j)$ is given one or zero for similar or dissimilar pairs, respectively. Similar pair indicates that they belong to the same category, and vice-versa. The idea is to minimize the distance between similar pairs and maximize the distance between dissimilar pairs. To this end, we employ the pairwise constraints *w.r.t* KL-divergence [167], since KL-divergence measures the discrepancy of two probability distributions and is non-negative. For a pair $(\mathbf{x}_i, \mathbf{x}_j)$, we obtain their corresponding probabilities of outputs as (s_i, s_j) via the inference of our model $s_i = f(g(\mathbf{x}_i))$. If the pair is similar, the objective is to draw them closer. As KL-divergence is asymmetric, the similarity loss can be written as:

$$\mathcal{L}^1(\mathbf{x}_i, \mathbf{x}_j) = \mathcal{D}_{KL}(s_i || s_j) + \mathcal{D}_{KL}(s_j || s_i), \quad (8.2)$$

$$\mathcal{D}_{KL}(s_i || s_j) = \sum_{k=1}^K s_i^k \log \frac{s_i^k}{s_j^k}, \quad (8.3)$$

where $\mathcal{D}_{KL}$ denotes the KL-divergence and s_i^k denotes the k -th value of the output probability vector s_i . In the reverse case, if the pair is dissimilar, the objective is to push two samples away by a margin m , which can be formulated by a hinge loss:

$$\mathcal{L}^0(\mathbf{x}_i, \mathbf{x}_j) = [m - \mathcal{D}_{KL}(s_i || s_j)]^+ + [m - \mathcal{D}_{KL}(s_j || s_i)]^+, \quad (8.4)$$

where $[\cdot]^+$ denotes the function $\max(0, \cdot)$. Then for any pair and its similarity label κ , we combine them together into a contrastive loss:

$$\mathcal{L}_{sim}(\mathbf{x}_i, \mathbf{x}_j) = \kappa \mathcal{L}^1(\mathbf{x}_i, \mathbf{x}_j) + (1 - \kappa) \mathcal{L}^0(\mathbf{x}_i, \mathbf{x}_j). \quad (8.5)$$

We formulate this objective as cluster loss that can be directly added to the total loss without modifying the network. Intuitively, this cluster loss considers both intra-class and inter-class information for feature learning. In CDM, we apply the cluster loss to the samples in the source domain and the target domain. As labels are not available in the target domain, similarity label κ is inferred by current network predictions $\hat{\mathcal{Y}}_t$. However, as the pseudo-label $\hat{y}_t$ is possibly wrong, our contrastive loss may iteratively draw dissimilar pairs closer. To alleviate such deleterious effect, we propose a weighted contrastive loss according to the current predictions for the cross-domain similarity learning. Denote z_i as the largest logit of the network output (after softmax function) given a sample $\mathbf{x}_i$. For a mini-batch with size N_b , we formulate the objective of weighted cluster loss as:

$$\mathcal{L}_c(\mathcal{X}_s, \mathcal{X}_t, \kappa(\mathcal{Y}_s, \hat{\mathcal{Y}}_t); \theta) = \frac{1}{N_b} \sum_{\forall i, j \in N_s} w_{ij} \mathcal{L}_{sim}(\mathbf{x}_i, \mathbf{x}_j), \quad (8.6)$$

$$w_{ij} = \begin{cases} 1 & \mathbf{x}_i, \mathbf{x}_j \in \mathcal{X}_s \\ z_i & \mathbf{x}_i \in \mathcal{X}_t, \mathbf{x}_j \in \mathcal{X}_s \\ z_j & \mathbf{x}_i \in \mathcal{X}_s, \mathbf{x}_j \in \mathcal{X}_t \\ z_i z_j & \mathbf{x}_i, \mathbf{x}_j \in \mathcal{X}_t \end{cases} \quad (8.7)$$

which forms tight clusters for the two domains and draws similar samples closer. The weight w_{ij} encourages high-confidence samples in the target domain to approach their similar samples while weakening the side effects of low-confidence samples. Using the weighted cluster loss, our model learns tight and discriminative clusters in the cross-domain setting.

8.2.3 Cluster-level Discrepancy Minimization

Having learned the tight and discriminative clusters, we propose to minimize the cluster-level discrepancy. Such discrepancy is mitigated from two perspectives. The first objective is to reduce the feature-level discrepancy regardless of the impact of

sample-level label shift. The second objective is to minimize the marginal distributions across prototypes of two domains. To this end, we first calculate the centroid (or prototype) [44] of cluster for each category. The centroid of the k -th category μ_k^s is obtained by calculating the mean embedding of samples in the source domain that belong to the k -th category:

$$\mu_k^s = \frac{1}{|N_k^s|} \sum_{\mathbf{x}_i^s \in \mathcal{X}_k^s} g(\mathbf{x}_i^s), \quad (8.8)$$

where N_k^s is the sample number of subset $\mathcal{X}_k^s$ that constitutes the samples from the k -th category in the source domain $\mathcal{X}_s$. Similarly, we assign pseudo labels to the samples by current network predictions in the target domain, and also obtain K centroids.

Having $2K$ centroids for all the categories in the two domains, we encourage the cluster in the target domain to approach their corresponding cluster in the source domain. The feature-level alignment renders the k -th centroid in the target domain to possess similar characteristics as the corresponding one in the source domain. The centroid alignment loss is defined as

$$\mathcal{L}_p(\mathcal{X}_s, \mathcal{X}_t; \theta) = \sum_{i=1}^K \|\mu_i^s - \mu_i^t\|^2, \quad (8.9)$$

where $\|\cdot\|$ is the norm. From the perspective of distribution, since there is no label shift between centroids of the two domains, *i.e.* $p(y|\mu) = q(y|\mu)$, the assumption of UDA holds for μ , which motivates us to align their conditional distributions by mapping them to RKHS and minimizing MMD. One drawback of applying MMD for UDA is that its alignment is irrelevant to the category, which can be overcome by the category-wise cluster alignment in Eq.(8.9). The distribution discrepancy loss is formulated as:

$$\mathcal{L}_d(\mathcal{X}_s, \mathcal{X}_t; \theta) = \frac{1}{K} \left\| \sum_{k=1}^K \phi(\mu_k^s) - \sum_{k=1}^K \phi(\mu_k^t) \right\|_{\mathcal{H}}^2, \quad (8.10)$$

where $\phi(\cdot)$ is the mapping to RKHS $\mathcal{H}$. In CDM, we use the optimal multi-kernel selection method [3].

The cluster alignment has been ever exploited in MSTN [77] and CAT [168]. However, the assumption and design of our method are different. MSTN fails to form

tight clusters before alignment. CAT learns the clusters using L2-distance of feature embeddings obtained by a teacher network, which ignores the similarities between categories and involves greater complexity. In comparison, the proposed CDM learns tight clusters via similarity-based learning, and takes both feature-level discrepancy ($\mathcal{L}_p$) and distribution-level discrepancy ($\mathcal{L}_d$) into account, which reduces the cross-domain shift regardless of the effects of label shift, thus leading to a robust domain-invariant representation learning.

8.2.4 Boundary Refinement

With the aligned and tight clusters derived from the first two objectives, we further propose to revamp the classifier such that the classifier has a reasonable boundary. Naturally, the decision boundary of classifier should be kept far away from data-dense regions in the feature space, which is called the cluster assumption [166]. This could be achieved via minimization of conditional entropy in terms of the target distribution:

$$\mathcal{L}_{ce}(\mathcal{X}_t; \theta) = -\mathbb{E}_{(\mathbf{x}, y) \sim \mathbb{Q}}[h_\theta(\mathbf{x})^T \ln h_\theta(\mathbf{x})]. \quad (8.11)$$

It has shown promising performance for semi-supervised learning and domain adaptation [169]. However, we argue that *boundary refinement is not sufficient without aligned and tight clusters*. Specifically, when clusters of one category in the source domain and the target domain are separate, the conditional entropy loss can wrongly push them away, which leads to negative transfer. In IDA, we incorporate boundary refinement into cluster learning and alignment, the combination of which enforces the boundary to occur in the inter-class region.

In summary, the overall objective is to minimize the weighted summation of these losses, formulated as:

$$\mathcal{L}_{total} = \mathcal{L}_y + \beta \mathcal{L}_c + \gamma(\mathcal{L}_p + \mathcal{L}_d) + \eta \mathcal{L}_{ce}, \quad (8.12)$$

where hyper-parameters $\{\beta, \gamma, \eta\}$ control the interactions of loss items. The overall objective is optimized via back-propagation.

8.2.5 Theoretical Insight

Denote f_S and $f_{\hat{T}}$ as the source labeling function and the target pseudo labeling function, respectively. Let h denote a hypothesis. The error that CDM minimizes is formulated as

$$\epsilon_\xi(h) = \xi\epsilon_t(h, f_{\hat{T}}) + (1 - \xi)\epsilon_s(h, f_S), \quad (8.13)$$

where ξ controls the interactions of supervised learning in Eq.(8.1) and the pseudo label learning in Eq.(8.4). Following the learning theory of domain adaptation [48], the terms $\epsilon_t(h, f_{\hat{T}})$ and $\epsilon_s(h, f_S)$ represent the expected error on the target domain and the source domain *w.r.t.* pseudo labels and ground truth labels. We can further obtain the upper bound of the difference between our objective $\epsilon_\xi(h)$ and the expected error on the target domain $\epsilon_t(h, f_T)$, which is derived by [78]. The less the difference is, the more effectively CDM performs.

Theorem 8.2.2. [78]

Let h be a hypothesis in a class $\mathcal{H}$. Then

$$|\epsilon_\xi(h) - \epsilon_t(h, f_T)| \leq (1 - \xi)\left(\frac{1}{2}d_{\mathcal{H}\Delta\mathcal{H}}(\mathcal{D}_s, \mathcal{D}_t) + \lambda\right) + \xi\epsilon_t(f_T, f_{\hat{T}}), \quad (8.14)$$

where $d_{\mathcal{H}\Delta\mathcal{H}}(\mathcal{D}_s, \mathcal{D}_t)$ measures the domain discrepancy in the hypothesis space $\mathcal{H}$, $\epsilon_t(f_T, f_{\hat{T}})$ is the rate of false pseudo labels, and λ is the minimal risk of the joint ideal hypothesis h^* defined as

$$\lambda = \min_{h \in \mathcal{H}} \epsilon_s(h^*, f_S) + \epsilon_t(h^*, f_T) \quad (8.15)$$

We show how CDM minimizes the three terms of the difference in Theorem 8.2.2.

(1) The first term is the domain discrepancy measure $d_{\mathcal{H}\Delta\mathcal{H}}(\mathcal{D}_s, \mathcal{D}_t)$ that is minimized by reducing the category-level MMD in Eq.(8.10), which has been proved in [170]. (2) In many methods [78, 170], the risk λ is considered to be negligibly small and thus disregarded. In IDA, the risk becomes larger than that of standard UDA due to label shift. We then apply the triangle inequality for classification error by

the learning theory of domain adaptation [48, 171].

$$\begin{aligned}
\lambda &= \min_{h \in \mathcal{H}} \epsilon_S(h, f_S) + \epsilon_T(h, f_T) \\
&\leq \min_{h \in \mathcal{H}} \epsilon_S(h, f_S) + \epsilon_T(h, f_S) + \epsilon_T(f_S, f_T) \\
&\leq \min_{h \in \mathcal{H}} \epsilon_S(h, f_S) + \epsilon_T(h, f_S) + \epsilon_T(f_S, f_{\hat{T}}) + \epsilon_T(f_T, f_{\hat{T}})
\end{aligned} \tag{8.16}$$

The first two terms represent the disagreement between h and f_S which are assumed to be fixed with source labels available. Considering the third term $\epsilon_T(f_S, f_{\hat{T}})$, as described in Eq.(8.9), CDM aligns the centroids obtained by the average of source samples and the pseudo-labeled target samples for the same class, which is formulated as

$$\epsilon_T(f_S, f_{\hat{T}}) = \mathbb{E}_{x \sim T} [|f_S(g(x)) - f_{\hat{T}}(g(x))|].$$

Therefore $\epsilon_T(f_S, f_{\hat{T}})$ should also be small. The last term of λ is the false pseudo label rate. (3) Both the third term of the bound in Theorem 8.2.2 and the last term of λ are the false pseudo label rate $\epsilon_T(f_T, f_{\hat{T}})$ which is reduced by progressive learning in CDM. We will empirically show how this rate decreases in Section 8.3.3. As discussed above, the bound in Eq.(8.14) can be constantly tightened.

8.3 Experiment

To demonstrate the efficacy of CDM in tackling the IDA problem, two protocols are firstly proposed to construct label shift on existing balanced UDA datasets, and three challenging benchmarks are employed including *Office-31*, *Image-CLEF* and *Office-Home*. Then we conduct ablation study to investigate the effect of label shift and visualize the aligned features.

8.3.1 Setup

Dataset Protocol. We construct the datasets with label shift by applying Pareto distribution ($\alpha = 100$) [162] to sub-sampling the existing UDA datasets. In this fashion, we design two protocols (*source domain* $\rightarrow$ *target domain*) for different objectives: (1) **Balanced** $\rightarrow$ **Unbalanced Protocol (B2U)** mimics practical situations where many public datasets are balanced, but the visual categories in

FIGURE 8.3: Dataset protocols for evaluation and real-world dataset SVHN.

reality are usually long-tailed. (2) **Unbalanced $\rightarrow$ inverse Unbalanced Protocol (U2U)** is designed for the extreme IDA situation where the label distribution in the target domain is an inverse order of that in the source domain. To circumvent few-shot learning that is irrelevant to UDA, we set the minimum number of samples for a category to be ten. Fig. 8.3 shows the distributions of two protocols that simulate the real-world distribution such as the Street View House Numbers (SVHN) Dataset [172].

To prove the robustness of CDM, we select three datasets with different characteristics. **Office-31** [94] is the most widely-used benchmark for UDA with 31 categories from three domains: *Amazon* (**A**), *Webcam* (**W**) and *DSLR* (**D**). As *DSLR* is too small (498 images), we only employ *Amazon* and *Webcam* for sub-sampling, forming two tasks **A** $\rightarrow$ **W** and **W** $\rightarrow$ **A**. Its difficulty lies in its imbalanced category and limited samples. **Image-CLEF** is a standard UDA benchmark with 12 categories from three domains including *Caltech-256* (**C**), *ImageNet ILSVRC 2012* (**I**) and *Pascal VOC 2012* (**P**). Each category in one domain contains 50 images, so we can apply our protocols to all three domains. **Office-Home** [95] is a very challenging UDA dataset since it consists of four domains including *Artistic images* (**Ar**), *Clipart images* (**Cl**), *Product images* (**Pr**) and *Real-World images* (**Rw**). As the sample number of *Artistic images* is small, we only employ the other three domains to construct 6 transfer tasks. With label shift in the target domain, evaluation on this dataset becomes the most challenging benchmark. More introductions on these benchmarks have been illustrated in Chapter 2.

Baselines. We follow standard evaluation protocols of UDA that utilize all labeled source data and all unlabeled target data for training. The evaluation criterion is the mean accuracy (with standard deviation) across three runs. We choose

AlexNet [146] and **ResNet-18** [141] as two backbones for evaluation. We compare our method with the state-of-the-art transfer learning methods: Deep Domain Confusion (**DDC**) [58], Deep Adaptation Network (**DAN**) [3], weighted Domain Adaptation Network (**WDAN**) [85], Deep Correlation Alignment (**CORAL**) [61], Domain Adversarial Neural Network (**DANN**) [75], Conditional Adversarial Network (**CDAN**) [145], Maximum Classifier Discrepancy (**MCD**) [73], Moving Semantic Transfer Network (**MSTN**) [77], Partial Adversarial Domain Adaptation (**PADA**) [87], Transferrable Prototype Network (**TPN**) [78] and Cluster Alignment with a Teacher (**CAT**) [168]. Among them, **DDC**, **DAN**, **WDAN** and **CORAL** aligns the distribution by some measure of discrepancy. **DANN**, **CDAN** and **MCD** are novel adversarial domain adaptation methods. **MSTN**, **TPN** and **CAT** conduct semantic alignment by pseudo labels. **PADA** addresses a specific label shift where the target domain is a subset of source domain. For a fair comparison, all the evaluations on these baselines (totally 508 settings) were conducted using their **official codes and hyper-parameters** for three datasets when available.

Implementation Details. We implement our method in **PyTorch** and employ the official ImageNet pre-trained parameters. For all experiments, we set $\lambda = 1, \gamma = 0.1, \eta = 0.01$ which are obtained by the empirical validation on *Office-31*. For MMD, the same settings as DAN [3] are used. For the cluster learning, the batch size of the source domain and the target domain are 32 and 128, respectively. We use the mini-batch stochastic gradient descent (SGD) with a momentum of 0.9 and decaying learning rate from 0.01 to 0.001 in [75]. For the feature extractor, $\frac{1}{10}$ of the standard learning rate is specified during training.

8.3.2 Summary of Results

The classification results of imbalanced domain adaptation on the two tasks on *Office-31*, six tasks on *Image-CLEF* and six tasks on *Office-Home* are shown in Table 8.1, 8.2 and 8.3, respectively. We record the average accuracy and standard deviation of three runs for all the methods. The proposed CDM outperforms the state-of-the-art methods based on two backbones (AlexNet and ResNet-18) and two protocols (B2U and U2U). It can be observed in Table 8.1 that based on ResNet-18 our method significantly improves the source-only model by 6.6% and 8.5% for

Method	A→W		W→A		Avg
AlexNet	51.0±1.3	44.7±0.9	51.7±1.2	46.5±1.2	51.4 / 45.6
DANN	35.5±4.5	35.2±3.2	44.6±1.5	45.6±1.3	40.1 / 40.4
CDAN	52.1±3.3	39.5±5.1	39.9±2.1	34.5±4.6	46.0 / 37.0
CORAL	53.6±4.5	45.2±3.3	49.7±0.0	47.2±1.9	51.7 / 46.2
DAN	54.5±3.1	44.5±0.0	49.9±1.5	47.0±1.4	52.2 / 45.8
CAT	55.3±1.7	42.5±1.0	49.3±1.2	47.2±1.1	52.3 / 44.9
MSTN	56.2±5.3	42.7±1.2	48.4±1.3	46.3±1.7	52.3 / 44.5
WDAN	55.5±1.3	45.1±1.0	49.5±1.2	47.3±1.5	52.5 / 46.2
PADA	58.3±4.7	45.0±3.0	46.8±4.3	45.2±2.8	52.6 / 45.1
MCD	56.7±0.7	48.4±5.1	51.3±0.6	46.9±2.5	54.0 / 47.7
CDM	60.0±0.2	52.3±0.4	52.5±0.5	50.0±0.2	56.3 / 51.2

ResNet18	73.4±1.1	61.0±1.0	64.2±0.6	56.7±1.7	68.8 / 58.9
MCD	69.0±1.8	66.6±1.6	60.0±0.9	56.4±0.0	64.5 / 61.5
CDAN	71.0±2.9	64.1±4.3	60.4±1.4	57.7±3.1	65.7 / 60.9
DDC	73.8±1.7	64.0±2.8	62.9±0.5	56.5±0.0	68.4 / 60.3
PADA	78.6±3.4	62.1±5.2	59.4±7.4	53.1±1.8	69.0 / 57.6
DAN	73.6±0.0	63.0±1.6	64.9±1.5	57.9±1.0	69.3 / 60.5
CORAL	73.8±1.2	62.8±0.6	64.9±0.0	58.7±0.4	69.4 / 60.8
DANN	74.8±2.2	65.2±1.7	65.0±0.4	54.7±0.0	69.9 / 60.0
CDM	81.3±0.4	70.9±0.6	69.4±0.5	62.2±0.6	75.4 / 66.6

TABLE 8.1: The average classification accuracy (%) with the standard deviation on *Office-31*. For each column, left and right results are obtained by applying **B2U** and **U2U** protocol, respectively.

FIGURE 8.4: The confusion matrix on A→W using U2U protocol.

B2U and U2U, respectively. In Table 8.2, CDM shows its superiority especially on the U2U protocol with larger label shift. We can see that for the AlexNet-based evaluation on *Image-CLEF*, only DAN, CORAL and CDM have improvement and CDM has smaller standard deviation, which indicates that our method achieves higher stability with label shift. In Table 8.3, CDM achieves the state-of-the-art performance for most tasks on *Office-Home* that has quite distinct domains. It demonstrates that our method can cope with the challenging covariate shift and label shift simultaneously. As for two backbones, though the initial rate of false pseudo labels for AlexNet is high, CDM still constantly improves the performance. The improvement of ResNet-18 is larger than that of AlexNet, which shows that better representation learning benefits the progressive learning in CDM.

We notice that all other methods encounter the problem of **negative transfer** to some extent. For instance, in Table 8.3, based on ResNet-18, MCD incurs performance drops for all the tasks of *Office-Home*, while for standard UDA, MCD has competitive accuracy among these approaches [73]. The most severe case is the evaluation on *Image-CLEF* based on ResNet-18 and U2U protocol. As shown in Table 8.2, all the methods except CDM suffer from performance degradation. PADA [87], which is designed to deal with the partial label shift problem, shows

Method	I→P	P→I	I→C	C→I	C→P	P→C	Avg
AlexNet	61.2±1.7 / 56.1±1.0	71.6±0.9 / 58.8±3.0	76.7±0.6 / 70.1±1.2	70.5±0.1 / 67.8±0.2	58.5±0.8 / 52.2±1.4	76.7±2.1 / 70.5±2.1	69.2 / 62.6
CDAN	55.2±0.0 / 42.4±6.0	67.2±6.3 / 43.3±2.4	70.1±2.3 / 57.0±3.0	58.5±5.4 / 47.2±4.5	46.9±0.6 / 35.8±2.7	64.8±6.0 / 40.0±2.7	60.4 / 44.3
MSTN	62.9±3.1 / 46.3±1.5	67.2±3.1 / 48.3±1.2	72.1±3.2 / 61.2±3.1	68.1±2.5 / 62.2±1.3	53.0±2.7 / 51.1±3.1	53.2±2.1 / 41.8±0.9	62.8 / 51.8
DANN	62.1±6.9 / 47.8±1.8	66.3±6.3 / 46.6±0.9	73.1±5.9 / 57.3±6.6	67.9±3.4 / 62.1±5.3	53.1±8.6 / 50.1±11.3	54.8±2.3 / 42.4±0.0	62.9 / 51.1
CAT	65.1±2.8 / 47.8±1.8	66.5±3.5 / 48.1±1.2	74.5±3.1 / 56.3±2.1	66.5±2.1 / 62.1±2.5	52.1±1.3 / 51.2±1.9	55.9±3.1 / 44.9±0.8	63.4 / 51.7
MCD	60.9±4.8 / 56.4±3.6	70.1±0.0 / 56.1±1.2	79.4±3.6 / 68.4±2.4	69.1±3.1 / 63.0±0.9	55.2±3.9 / 51.6±1.2	75.2±3.6 / 62.1±1.5	68.4 / 59.6
PADA	62.7±1.8 / 43.3±2.3	72.2±1.5 / 48.1±6.3	81.6±0.6 / 72.0±4.7	60.7±1.0 / 54.6±3.1	55.0±0.4 / 47.8±2.9	80.1±0.7 / 49.0±2.0	68.7 / 52.5
TPN	61.3±3.3 / 56.5±3.1	70.5±0.5 / 56.7±1.5	80.1±2.8 / 68.2±2.1	68.9±2.5 / 62.8±0.9	56.5±1.9 / 53.1±1.5	75.5±2.3 / 62.3±1.7	68.8 / 59.9
CORAL	63.6±3.9 / 59.7±3.9	69.9±3.0 / 62.1±7.2	77.3±2.1 / 72.5±2.9	70.7±0.3 / 69.0±1.5	57.7±1.0 / 57.0±0.9	79.5±1.3 / 71.0±5.0	69.8 / 65.2
DAN	63.0±1.2 / 57.4±3.7	71.9±0.6 / 58.8±1.8	78.0±1.0 / 73.2±2.8	72.8±1.2 / 66.9±2.1	58.2±0.9 / 56.1±1.2	77.9±1.2 / 72.8±2.1	70.3 / 64.2
WDAN	63.2±1.5 / 58.1±2.1	72.1±1.4 / 58.5±1.2	78.3±1.0 / 73.2±2.1	71.9±1.4 / 67.1±2.3	59.1±1.1 / 56.3±1.0	78.1±0.9 / 72.8±1.3	70.5 / 64.3
CDM	64.5±0.5 / 59.3±0.7	72.5±0.5 / 65.4±2.2	81.7±0.2 / 76.1±0.9	71.3±0.0 / 68.7±0.1	59.8±0.7 / 57.8±0.5	80.9±0.5 / 73.1±0.0	71.8 / 66.7
ResNet	73.1±0.5 / 65.6±1.1	85.6±1.1 / 75.8±0.0	89.5±0.3 / 77.9±0.9	85.3±0.9 / 83.3±0.5	71.3±1.5 / 67.5±0.0	87.1±1.4 / 76.4±0.9	82.0 / 74.4
CORAL	72.5±0.0 / 67.5±0.3	86.9±2.1 / 75.5±2.4	87.8±0.5 / 76.4±1.5	85.1±0.6 / 80.6±0.9	70.1±1.7 / 65.1±0.6	87.5±1.2 / 74.6±3.6	76.7 / 73.3
CDAN	70.8±1.5 / 67.8±4.8	82.7±6.3 / 67.2±9.9	86.6±4.8 / 78.5±8.9	78.2±6.9 / 73.1±8.3	66.0±5.1 / 62.4±8.1	79.4±8.4 / 68.9±6.2	77.3 / 70.0
DDC	72.5±0.6 / 66.0±2.1	84.5±1.8 / 74.0±2.4	88.1±0.1 / 77.0±0.0	87.2±2.1 / 79.7±0.9	69.3±0.3 / 63.9±0.0	87.2±3.3 / 71.6±0.0	81.5 / 72.0
MCD	73.4±1.5 / 66.9±0.0	82.1±0.0 / 66.0±0.9	91.3±3.2 / 74.3±1.8	88.4±3.6 / 83.3±1.8	70.4±2.0 / 67.8±0.0	85.7±4.2 / 62.1±0.0	81.9 / 70.1
DAN	72.5±0.0 / 67.5±0.0	87.5±1.2 / 71.6±2.0	89.9±1.2 / 80.0±0.0	86.3±2.4 / 78.8±0.3	69.9±3.0 / 66.0±0.9	87.2±1.5 / 76.1±0.0	82.2 / 73.3
DANN	73.4±2.1 / 65.7±0.6	87.8±3.0 / 72.2±0.0	88.7±1.2 / 76.7±0.6	86.3±0.0 / 80.3±1.5	71.0±1.6 / 65.7±0.9	86.6±1.5 / 72.2±0.3	82.3 / 72.1
PADA	71.0±1.4 / 65.9±4.9	86.9±2.4 / 63.9±3.2	92.2±0.9 / 79.4±4.8	88.0±2.6 / 75.5±5.5	71.9±0.6 / 60.6±4.6	86.3±0.9 / 76.8±4.9	82.7 / 70.4
CDM	74.3±0.8 / 69.3±0.7	91.6±1.2 / 76.7±0.0	91.4±0.5 / 83.0±1.4	90.2±0.6 / 85.4±0.0	71.3±0.8 / 70.2±0.3	90.2±1.2 / 77.1±0.0	84.8 / 77.0

TABLE 8.2: The average classification accuracy (%) with the standard deviation on *Image-CLEF*. For each column, left and right results are obtained by applying **B2U** and **U2U** protocol, respectively.

Method	Cl→Pr	Cl→Rw	Pr→Cl	Pr→Rw	Rw→Cl	Rw→Pr	Avg
AlexNet	35.1±0.2 / 25.1±0.4	39.5±0.0 / 27.5±0.2	28.3±0.3 / 21.5±0.3	48.6±0.3 / 34.3±0.2	34.1±0.8 / 30.9±0.4	57.5±0.4 / 42.6±0.5	40.5 / 30.3
DANN	31.7±1.5 / 21.4±0.4	33.6±0.1 / 23.1±3.2	25.2±0.9 / 20.7±1.3	45.9±1.3 / 32.2±1.4	27.1±1.9 / 24.0±0.7	54.4±0.7 / 40.4±0.9	36.3 / 27.0
PADA	33.6±10.4 / 24.5±7.0	37.5±7.0 / 27.7±5.4	26.6±3.6 / 19.2±7.1	46.5±3.2 / 34.5±8.2	32.3±2.7 / 25.9±9.9	55.1±2.3 / 42.7±4.1	38.6 / 29.1
CDAN	35.1±3.4 / 26.0±3.6	38.3±3.0 / 23.8±5.1	28.4±1.3 / 20.3±2.2	45.5±1.7 / 33.5±1.0	36.3±2.3 / 30.7±3.4	56.7±1.5 / 43.1±4.4	40.1 / 29.6
CORAL	37.0±1.9 / 28.4±1.5	41.1±2.2 / 30.6±0.4	30.0±0.5 / 23.9±0.5	50.3±1.6 / 40.4±0.0	34.9±0.7 / 30.9±0.0	58.7±0.2 / 49.4±0.5	42.0 / 34.0
DAN	36.8±1.8 / 30.1±0.0	41.1±0.2 / 30.9±0.4	31.5±0.8 / 24.8±0.6	51.1±1.4 / 40.2±0.5	37.9±0.0 / 32.0±0.0	58.4±1.8 / 48.2±0.5	42.8 / 34.4
MCD	37.2±1.0 / 29.8±1.2	42.8±0.4 / 31.8±0.8	31.3±2.0 / 25.8±0.1	50.5±0.6 / 38.8±1.1	37.0±1.3 / 34.3±0.0	58.3±1.1 / 46.8±1.2	42.9 / 34.6
CDM	41.1±0.3 / 32.5±0.5	43.9±1.1 / 33.8±0.1	31.6±0.5 / 25.4±0.4	51.4±0.3 / 42.0±0.3	38.1±1.2 / 34.6±0.1	59.8±0.9 / 48.0±0.2	44.3 / 36.1
ResNet	54.7±0.8 / 43.3±0.0	58.1±0.5 / 44.5±0.4	38.0±1.0 / 30.8±1.0	66.2±0.5 / 56.4±1.0	43.9±0.5 / 39.7±0.9	72.8±0.2 / 66.7±0.7	55.8 / 46.9
PADA	22.5±0.3 / 14.8±0.8	46.3±3.1 / 30.2±4.1	28.7±0.0 / 17.2±0.6	54.8±0.4 / 33.4±0.0	42.2±0.8 / 33.6±6.5	70.2±0.2 / 57.4±7.1	44.1 / 31.1
MCD	49.2±1.2 / 37.5±1.2	53.2±0.0 / 38.4±0.0	35.6±0.2 / 30.7±0.0	64.6±0.0 / 53.8±1.7	41.3±0.5 / 36.7±0.0	66.8±0.0 / 59.7±0.1	51.8 / 42.8
CDAN	50.6±1.8 / 39.6±2.9	55.3±1.9 / 42.3±3.3	36.4±1.1 / 27.0±2.4	62.8±3.5 / 50.1±3.6	46.9±1.1 / 37.6±1.5	72.1±0.4 / 62.8±3.2	54.0 / 43.2
DANN	54.4±0.5 / 42.6±0.2	58.5±0.9 / 47.8±0.0	41.3±0.9 / 31.7±0.0	67.0±0.1 / 55.8±0.1	46.7±0.6 / 41.9±0.4	71.1±0.6 / 64.0±0.9	56.5 / 47.3
DDC	54.9±0.7 / 41.9±0.0	57.9±1.2 / 46.5±0.9	40.4±0.4 / 32.0±0.0	67.4±0.2 / 54.0±0.4	48.3±0.7 / 42.2±0.6	70.9±0.6 / 62.2±0.2	56.6 / 46.5
CORAL	54.8±0.9 / 42.6±0.5	59.9±0.7 / 47.2±0.6	41.2±0.3 / 33.0±0.8	66.9±0.4 / 56.0±0.0	47.6±0.6 / 42.8±1.2	71.2±0.7 / 64.0±0.6	56.9 / 47.6
DAN	54.5±0.6 / 43.6±0.8	59.7±0.7 / 47.5±0.7	41.0±0.7 / 32.5±0.2	67.8±0.5 / 55.4±0.5	48.2±0.3 / 41.5±0.8	71.5±0.4 / 64.2±0.1	57.1 / 47.5
CDM	57.4±0.6 / 45.3±0.8	63.8±0.1 / 52.4±0.1	44.4±0.1 / 33.6±0.3	68.6±0.2 / 58.5±0.6	50.8±0.3 / 43.9±0.7	73.7±0.2 / 67.2±0.5	59.8 / 50.2

TABLE 8.3: The average classification accuracy (%) with the standard deviation on *Office-Home*. For each column, left and right results are obtained by applying **B2U** and **U2U** protocol, respectively.

some improvements on *Office-31* and *Image-CLEF*. However, it fails when confronting large covariate shift and label shift on *Office-Home*. The cluster-based MSTN and CAT also suffer from negative transfer to some extent. These results show that when label shift is large, aligning the marginal distribution cannot improve but hurt the performance in the target domain [164].

8.3.3 Analysis

Ablation Study. We next validate the effectiveness of three parts: cluster learning ($\mathcal{L}_c$), cluster alignment ($\mathcal{L}_p + \mathcal{L}_d$) and boundary refinement ($\mathcal{L}_{ce}$). We choose four tasks with different levels of covariate shift from *Office-31* and *Office-Home* and conduct ablation study following B2U and U2U protocols based on ResNet-18. The results are shown in Table 8.4. The discriminative cluster learning leads to

Protocol Setting	B2U				U2U				Avg
	A→W	W→A	Cl→Rw	Rw→Cl	A→W	W→A	Cl→Rw	Rw→Cl	
Source-Only Model	73.4	62.7	58.1	43.9	57.6	56.7	44.5	39.7	54.6
Cluster Learning	76.7	67.9	61.1	49.4	68.1	60.5	50.4	41.7	59.5
Cluster Alignment	79.3	68.7	62.5	50.8	69.7	62.2	51.0	43.1	60.9
Boundary Refinement	81.3	69.4	63.8	50.5	70.9	62.6	52.4	43.9	61.9

TABLE 8.4: The classification accuracy (%) for ablation study on *Office-31* and *Office-Home* following B2U and U2U protocols.

FIGURE 8.5: (a) The accuracy on A→W with different level of label shift. (b)(c) The comparison of average accuracy on *Image-CLEF* based on ResNet18.

great improvement for all tasks, and the cluster alignment further improves the performance by aligning the cross-domain categories. As the boundary refinement keeps the classifier far away from clusters, it brings some improvements on seven tasks. For some cases such as $Rw \rightarrow Cl$, this condition has already been held and then the effect of the boundary refinement is not obvious.

Effect of Label Shift. As shown in Fig. 8.5(b)(c), the improvement margin of CDM is larger on U2U than that of B2U. Since B2U has less label shift than U2U, CDM performs more effectively for dealing with label shift. We also discover that label shift leads to larger training instability in terms of some adversarial adaptation approaches such as DANN, CDAN and PADA. In Table 8.1, we can see that PADA and DANN achieve accuracy improvement but their standard deviations are large. Similar results are observed in Table 8.2 and 8.3. We speculate that this is caused by the imbalanced training data that hinders adversarial training.

Class-level Accuracy. To further analyze whether CDM improves the accuracy of each category for the long-tailed test set, we show the confusion matrices of CDM and source-only model in Fig. 8.4. It is easily observed that label shift severely decreases the performance especially for the category that has imbalanced

FIGURE 8.6: The t-SNE visualization of DAN [3] and the proposed CDM on $P \rightarrow I$ (*Image-CLEF*). The left figure is with the domain label (red: source; blue: target) while the right figure is with the category label (in different colors).

data (*e.g.* 15-31 rows). CDM has better performance for most categories, which demonstrates the effectiveness of handling imbalanced classes.

Different Levels of Label Shift. As label shift increases the difficulty of domain adaptation in our setting, we quantitatively evaluate its effect by generating data splits with different levels of label shift. Evaluation is conducted on the task $A \rightarrow W$ (*Office-31*). We specify different α of the Pareto distribution in U2U protocol while keeping the total number of image samples invariant. When α increases, label shift of U2U protocol becomes smaller. Results are shown in Fig. 8.5(a), and we find that evaluations with larger α have better results. The Pareto distributions ($\alpha = 10, 50$) are so sharp that errors in the target domain increase tremendously. In reality, label shift also causes insufficient representation learning, *i.e.* few-shot learning, which can be solved by data-efficient deep learning models [173].

Feature Visualization. We visualize the t-SNE [147] on feature embeddings of DAN [3] and our model on $P \rightarrow I$ (*Image-CLEF*, B2U) in Fig. 8.6. Label shift occurs between the source domain and the target domain. Compared to DAN, the proposed CDM learns tight and discriminative clusters for classification. The rate

of false pseudo labels $\epsilon_T(f_T, f_{\hat{T}})$ is 46.9% after the first epoch and the final rate is about 9%, which conforms with our discussions in Section 8.2.5.

8.4 Conclusion

In this chapter, we introduced an imbalanced domain adaptation, a novel UDA scenario where label shift exists between the source domain and the target domain. To deal with this, we proposed a Cluster-level Discrepancy Minimization approach that reduces covariate shift and label shift concurrently. We theoretically and empirically deduced that directly aligning the marginal distribution of feature embeddings results in negative transfer. Extensive experiments were conducted to evaluate our method based on two protocols, and the results demonstrated that our method effectively transfers the knowledge and achieves the state-of-the-art performance on all benchmarks.

Chapter 9

Conclusion and Future Work

9.1 Conclusions

The proliferation of sensing devices and the development of machine learning algorithms propel the advent of smart homes and smart buildings, which significantly facilitates human lives and optimizes social order. This thesis focuses on robust sensing and recognition technologies by developing new sensing techniques and transfer learning algorithms. We first analyzed the challenges of current sensing technologies for smart homes and proposed a cost-effective, privacy-preserving and device-free WiFi-based sensing platform. Three statistical learning algorithms were designed and implemented using this platform for different applications including occupancy detection, activity recognition and gesture recognition. Since domain shift usually hinders the generalization of recognition models such as spatial shift in WiFi data and style difference in images, great efforts on developing robust domain adaptation algorithms have been made in this thesis. Four UDA methods have been proposed to tackle real-world adaptation problems. Both theoretical and practical contributions were made.

For WiFi-based sensing, the contributions are listed as follows:

- In smart homes, current sensing techniques still encounter some issues such as privacy concerns and insufficient granularity. To this end, Chapter 3 proposed an agile WiFi-based IoT-enabled sensing platform. It has been so far the first platform that contains all essentials of an IoT system, which can enable many

applications such as occupancy detection and human activity recognition in smart homes.

- Chapter 4 proposed a novel algorithm to enable a more difficult WiFi-based sensing application — human gesture recognition. A deep Siamese network architecture is proposed to extract useful and robust patterns from CSI data. The proposed method empowers few-shot or one-shot learning, which only leverages few samples in the environment to obtain a recognition model.

The thesis has made several contributions to the community of transfer learning for classification problem, which also benefits the recognition technology in smart sensing. The contributions are listed as follows:

- Domain shift is commonly confronted in visual and WiFi-based sensing. Great efforts on domain adaptation have been made to deal with it. However, as the computation resources are rather limited in the edge, the complexity of the recognition model is also constrained. Chapter 5 defined this problem as EDA and proposed MobileDA to tackle it. MobileDA inherits the advantages of UDA and model distillation, which is appealing for IoT applications.
- Chapter 6 deals with the stability of adversarial domain adaptation algorithms for automatic training in unmanned scenarios. In Chapter 6, we proposed MDAT that leverages a reconstruction network with margin loss to obtain stable gradients from the auxiliary domain classifier. The theoretical optimal solution is derived by the Nash Equilibrium, and the learning theories of domain adaptation are consistent with the objectives of MDAT. MDAT is a stable domain adaptation method that can be adopted in unmanned systems without cumbersome hyper-parameter tuning.
- Chapter 6 aims to stabilize the training of DAT, and Chapter 7 further explores how MDAT achieves better domain-invariant representations than DAT. To this end, spectral analysis is employed to quantify the transferability and discriminability of features learned by MDAT and DAT. Then we proposed AADA that generalizes MDAT to an asymmetric form. Theoretically, utilizing any energy functions as domain discriminator should work with AADA.

- Most UDA approaches are propitious to overcome the negative transfer by covariate shift, but ignores the label shift. Chapter 8 formulates the UDA problem with both covariate and label shift as IDA, which is commonly confronted such as long-tailed data that takes up the majority of real-world data. To this end, CDM was proposed in Chapter 8 which minimizes the discrepancy in cluster level and empowers domain adaptation for real-world imbalanced data.

9.2 Future Works

The statistical learning and transfer learning algorithms presented in this thesis advance robust sensing and recognition technologies. The successes of these methods motivate us to step up efforts to conduct in-depth research. Based on the research work in this thesis, there are some possible directions for future works:

- **Physical and statistical model for WiFi-based sensing** For WiFi-based sensing, CSI data is the source material and the proposed statistical learning algorithms are tools. Except for data-driven model, physical model describes how wireless signals propagate and their characteristics via closed-form formula. Fresnel zone theory was proposed for CSI and WiFi-based sensing [174], which can be employed for fall detection or even breath detection under ideal circumstances. Since real-world environments are complex and CSI are usually collected with noise, model-based method may not support complicated environments. Data-driven model has powerful fitting ability but lacks of physical understanding. Therefore, integrating both data-driven and model-based methods may enable robust and fine-grained WiFi-based sensing.
- **More tools and platforms for CSI** Though the proposed platform in Chapter 3 has equipped COTS WiFi router with CSI extraction, CSI extraction tools should be generalized to more devices. With the explosive proliferation of mobile devices and smart appliances [175], more wireless communication links are formed naturally without any extra costs. More CSI extraction tools for these devices provide more opportunities for comprehensive and accurate WiFi-based sensing.

- **Real-world challenges in domain adaptation** Numerous domain adaptation algorithms have been developed for covariate shift in domain adaptation problem following the standard protocol [2]. Nevertheless, in real-world scenarios, there are more challenges with respect to data and computational resources. For example, IDA and EDA in this thesis are two practical cases. More challenges include noisy annotations in the target domain and privacy-preserving issue in federated target domains.
- **Domain adaptation applications** Recent domain adaptation researches mainly pay attention to visual domain adaptation and sentiment classification [176]. Actually such covariate shift is commonly confronted in many applications such as the spatial shift in WiFi-based sensing. Another example is that in industrial IoT, UDA also works for fault diagnosis in machine health monitoring [177]. More UDA applications should be exploited with new domain-specific algorithms.

Apart from the aforementioned topics, we believe that the progresses of machine learning and IoT intensify artificial intelligence, thus benefiting human civilizations. In the future, the following research can significantly enhance the robustness of sensing and recognition system.

9.2.1 Machine Learning

Meta-Learning Traditional statistical learning demands large scales of training data that may not be available in practice. Meta-learning aims to train a model on a variety of learning tasks and then learn new tasks using a small number of samples. Model-Agnostic Meta-Learning (MAML) is a classic framework to solve classification and regression problems [178]. As the concept of meta-learning is more similar to the perceptual understanding of human beings, its research may contribute to the emergence and development of strong AI. Meta-learning algorithms can promote data-efficient sensing for some new scenarios.

Model interpretability The interpretability of deep learning severely hinders its availability in some critical fields such as medical and financial engineering [179]. Deep models are black-boxes that cannot provide intuitive representations and the key logic of judgment. Improved interpretation of the predictions will play a vital

role in applying deep learning to making critical decisions. For sensing systems in autonomous driving, good interpretability can help offer risk estimation that contributes to better decisions for smart vehicles. Moreover, if model interpretability becomes clear, reverse engineering will provide useful insights for many problems that cannot be currently modeled by physical theories.

9.2.2 Artificial Intelligence of Things (AIoT)

Generally, IoT refers to a system of interrelated everyday objects, usually with sensing or computing ability [180]. IoT is able to quantify daily lives of human beings or manufacturing of factories, which offers tremendous opportunities for many novel applications. Artificial Intelligence of Things (AIoT) is a combination of AI and IoT infrastructures, which leverages AI for more efficient IoT operations and obtaining critical decisions by IoT data. AIoT is the pivotal technology for smart homes and smart buildings. For example, in smart homes, intelligent personal assistant captures human orders by Amazon Echo [181] and then extracts the semantic information by AI. In smart buildings, many environment sensors are deployed so that AI algorithms assimilate these information and offer energy-saving strategy [182].

Multi-modal Learning In AIoT, the heterogeneous data is collected for a variety of sensors, which is called multi-modal data. As different sensor reflects different aspect of real-world situations, the connections among these modalities and the unique features of a specific modality are appealing for AIoT tasks. Multi-modal learning aims to extract robust representations from these heterogeneous data [183]. It is propitious to robust sensing and recognition in real-world applications.

Interdisciplinary Research Since IoT infrastructures offer sufficient data and AI algorithms own human-like capacity of achieving a specific target, AIoT begins to penetrate into various industries such as traffic engineering, pharmaceutical industry and civil engineering. Interdisciplinary research will be essential for the next wave of AIoT.

List of Author's Awards, Patents, and Publications¹

Journal Articles

(A) Journal papers that are included in the thesis

1. **J. Yang**, H. Zou, S. Cao, Z. Chen, and L. Xie, "MobileDA: Towards Edge Domain Adaptation," Accepted by *IEEE Internet of Things Journal*, 2020.
2. **J. Yang**, H. Zou, Y. Zhou, and L. Xie, "Learning Gestures From WiFi: A Siamese Recurrent Convolutional Architecture," *IEEE Internet of Things Journal*, vol. 6, no. 6, pp. 10763-10772, 2020.
3. **J. Yang**, H. Zou, H. Jiang, and L. Xie, "Device-free occupant activity sensing using WiFi-enabled IoT devices for smart homes," *IEEE Internet of Things Journal*, vol. 5, no. 5, pp. 3991-4002, 2018.
4. **J. Yang**, H. Zou, H. Jiang, and L. Xie, "CareFi: Sedentary behavior monitoring system via commodity WiFi infrastructures," *IEEE Transactions on Vehicular Technology*, vol. 67, no. 8, pp. 7620-7629, 2018.

(B) Journal papers that are not included in the thesis

1. **J. Yang**, H. Zou, Y. Zhou, and L. Xie, "Attention-based Domain-confused Adversarial Domain Adaptation for Real-world Cross-domain Object Recognition," Accepted by *Neurocomputing*, 2020

¹The superscript * indicates joint first authors

2. H. Zou, C. Chen, M. Li, **J. Yang**, Y. Zhou, L. Xie, and C. Spanos, "Adversarial Learning Enabled Automatic WiFi Indoor Radio Map Construction and Adaptation with Mobile Robot," Accepted by *IEEE Internet of Things Journal*, 2020.
3. K. Chen, **J. Yang**, J. Cheng, W. Chen, and C. Li, "Transfer learning enhanced AR spatial registration for facility maintenance management," *Automation in Construction*, vol. 113, no. 8, pp. 103-135, 2020.
4. K. Wang, X. Peng, **J. Yang**, D. Meng, and Y. Qiao, "Region attention networks for pose and occlusion robust facial expression recognition," *IEEE Transactions on Image Processing*, vol. 29, pp. 4057-4069, 2020.
5. W. Yang, Y. Yuan, ..., **J. Yang**, *et al.* "Advancing Image Understanding in Poor Visibility Environments: A Collective Benchmark Study," Accepted by *IEEE Transactions on Image Processing*, 2020.
6. Z. Chen, H. Zou, **J. Yang**, H. Jiang, and L. Xie, "WiFi fingerprinting indoor localization using local feature-based deep LSTM," Accepted by *IEEE Systems Journal*, 2019.
7. Y. Xu, **J. Yang**, and K. Mao, "Semantic-filtered Soft-Split-Aware video captioning with audio-augmented feature," *Neurocomputing*, vol. 357, pp. 24-35, 2019.
8. H. Zou, Y. Zhou, **J. Yang**, and C. Spanos, "Towards occupant activity driven smart buildings via WiFi-enabled IoT devices and deep learning," *Energy and Buildings*, vol. 177, pp. 12-22, 2018.
9. H. Zou, Y. Zhou, **J. Yang**, and C. Spanos, "Unsupervised wifi-enabled IoT device-user association for personalized location-based service," *IEEE Internet of Things Journal*, vol. 6, no. 1, pp. 1238-1245, 2018.
10. H. Zou, Y. Zhou, **J. Yang**, and C. Spanos, "Device-free occupancy detection and crowd counting in smart buildings with WiFi-enabled IoT," *Energy and Buildings*, vol. 174, pp. 309-322, 2018.
11. H. Zou, Y. Zhou, **J. Yang**, C. Spanos, and L. Xie, "Non-intrusive occupancy sensing in commercial buildings," *Energy and Buildings*, vol. 154, pp. 633-643, 2017.

12. L. Chen, Z. Li, and **J. Yang**, "Compressive perceptual hashing tracking," *Neurocomputing*, vol. 239, pp. 69-80, 2017.
13. Y. Xu, **J. Yang**, K. Mao, J. Yin, and S. See, "Exploiting Inter-Frame Regional Correlation for Efficient Action Recognition," *Under Review*.
14. Y. Xu, **J. Yang**, H. Cao, K. Mao, J. Yin, and S. See, "ARID: A New Dataset for Recognizing Action in the Dark," *Under Review*.
15. Y. Xu, H. Cao, **J. Yang**, K. Mao, J. Yin, and S. See, "PNL: Efficient Long-Range Dependencies Extraction with Pyramid Non-Local Module for Action Recognition," *Under Review*.
16. D. Wang, **J. Yang**, W. Cui, L. Xie, and S. Sun, "Multimodal CSI-based Human Activity Recognition using GANs," *Under Review*.
17. K. Chen, Z. Zeng, and **J. Yang**, "A Deep Region-based Pyramid Neural Network for Automatic Detection and Multi-Classification of Various Surface Defects of Aluminum Alloys," *Under Review*.
18. **J. Yang**, H. Qian, Han. Zou, and L. Xie, "Hierarchical Transfer Network for Unsupervised Domain Adaptation", *Under Review*.
19. J. Yang, L. Luo, W. Liao, L. Tang, S. Wang, **J. Yang**, Z. Yang, and H. Zhong, "Exploring and Improving Blurred Object Recognition with a Benchmark", *Under Review*.
20. **J. Yang**, Z. Zeng, K. Wang, H. Zou, and L. Xie, "GarbageNet: A Unified Learning Framework for Robust Garbage Classification", *Under Review*.
21. H. Qian, Y. Li, **J. Yang**, K. Tan, and L. Xie, "Deep Transfer Learning for Image-Based Microstructure Classification of Concrete Mortar and Paste Using Convolutional Neural Networks", *Under Review*.

Conference Proceedings

These conference papers in submission are included in this thesis.

1. **J. Yang**, H. Zou, Y. Zhou, Z. Zeng and L. Xie, "Asymmetric adversarial domain adaptation," Accepted by *European Conference on Computer Vision (ECCV)*, 2020.
2. **J. Yang**, H. Zou, Y. Zhou, and L. Xie, "Towards stable and comprehensive domain alignment: max-margin domain-adversarial training," *Under review*.
3. **J. Yang**, J. Yang, H. Zou, S. Cao, S. Wang, and L. Xie, "Cluster-level discrepancy minimization for imbalanced domain adaptation" *Under review*.
4. **J. Yang**, H. Zou, H. Jiang, and L. Xie, "Fine-grained adaptive location-independent activity recognition using commodity WiFi," in *IEEE Wireless Communications and Networking Conference (WCNC)*, 2018.

These conference papers are not included in this thesis.

1. S. Ji, K. Wang, X. Peng, **J. Yang**, Z. Zeng, and Y. Qiao, "Multiple Transfer Learning and Multi-Label Balanced Training Strategies for Facial AU Detection in the Wild," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, 2020.
2. **J. Yang**, K. Wang, X. Peng and Y. Qiao, "Deep recurrent multi-instance learning with spatio-temporal features for engagement intensity prediction," in *ACM International Conference on Multimodal Interaction*, pp. 594-598, 2018.
3. X. Peng, K. Wang, Z. Zeng, Q. Li, **J. Yang**, and Y. Qiao, "Suppressing Misabeled Data via Grouping and Self-Attention," in *European Conference on Computer Vision (ECCV)*, 2020.
4. K. Wang, X. Peng, **J. Yang**, S. Lu, and Y. Qiao, "Suppressing Uncertainties for Large-Scale Facial Expression Recognition," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020.
5. K. Wang*, **J. Yang***, D. Guo, K. Zhang, X. Peng and Y. Qiao, "Bootstrap Model Ensemble and Rank Loss for Engagement Intensity Regression," in *ACM International Conference on Multimodal Interaction*, pp. 551-556, 2020.

6. H. Chen, **J. Yang**, Y. Xu, Z. Shi, R. Luo, L. Xie, and R. Su, "Domain Adaptation for Degraded Remote Scene Classification" in *International Conference on Control, Automation, Robotics and Vision (ICARCV)*, 2020
7. D. Wang, **J. Yang**, W. Cui, L. Xie, and S. Sun, "Robust CSI-based Human Activity Recognition using Roaming Generator" in *International Conference on Control, Automation, Robotics and Vision (ICARCV)*, 2020
8. C. Wang, **J. Yang**, L. Xie, and J. Yuan, "Kervolutional Neural Networks," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019.
9. H. Zou, **J. Yang**, H. Das, H. Liu, Y. Zhou, and C. Spanos, "Wifi and vision multimodal learning for accurate and robust device-free human activity recognition," in *IEEE Conference on Computer Vision and Pattern Recognition Workshops*, 2019.
10. H. Zou, H. Das, **J. Yang**, Y. Zhou, and C. Spanos, "Machine Learning empowered Occupancy Sensing for Smart Buildings," in *IEEE International Conference on Machine Learning (ICML) Climate Change Workshop*, 2019.
11. H. Zou, **J. Yang**, H. Das, H. Liu, Y. Zhou, and C. Spanos, "Consensus Adversarial Domain Adaptation," in *The Thirty-Third AAAI Conference on Artificial Intelligence (AAAI-19)*, 2019.
12. H. Zou, **J. Yang**, Y. Zhou, and C. Spanos, "Joint Adversarial Domain Adaptation for Resilient WiFi-Enabled Device-Free Gesture Recognition," in *IEEE International Conference on Machine Learning and Applications (ICMLA)*, 2019.
13. K. Wang, X. Zeng, **J. Yang**, D. Meng, K. Zhang, X. Peng, and Y. Qiao, "Cascade attention networks for group emotion recognition with face, body and image cues," in *ACM International Conference on Multimodal Interaction*, pp. 640-645, 2018.
14. H. Zou, **J. Yang**, Y. Zhou, L. Xie, and C. Spanos, "Robust WiFi-enabled device-free gesture recognition via unsupervised adversarial domain adaptation," in *IEEE International Conference on Computer Communication and Networks (ICCCN)*, 2018.

15. H. Zou, Y. Zhou, **J. Yang**, H. Jiang, L. Xie, and C. Spanos, "WiFi-enabled device-free gesture recognition for smart home automation," in *IEEE 14th International Conference on Control and Automation (ICCA)*, 2018.
16. H. Zou, Y. Zhou, **J. Yang**, H. Jiang, L. Xie, and C. Spanos, "Deepsense: Device-free human activity recognition via autoencoder long-term recurrent convolutional network," in *IEEE International Conference on Communications (ICC)*, 2018.
17. H. Zou, Y. Zhou, **J. Yang**, W. Gu, L. Xie, and C. Spanos, "Wifi-based human identification via convex tensor shapelet learning," in *Thirty-Second AAAI Conference on Artificial Intelligence (AAAI)*, 2018.
18. H. Zou, Y. Zhou, **J. Yang**, W. Gu, L. Xie, and C. Spanos, "Multiple kernel representation learning for WiFi-based human activity recognition," in *IEEE International Conference on Machine Learning and Applications (ICMLA)*, 2017.
19. H. Zou, Y. Zhou, **J. Yang**, W. Gu, L. Xie, and C. Spanos, "Freecount: Device-free crowd counting with commodity wifi," in *IEEE Global Communications Conference (Globecom)*, 2017.
20. H. Zou, Y. Zhou, **J. Yang**, W. Gu, L. Xie, and C. Spanos, "Poster: WiFi-based device-free human activity recognition via automatic representation learning," in *Annual International Conference on Mobile Computing and Networking (Mobicom)*, 2017.
21. H. Zou, Y. Zhou, **J. Yang**, W. Gu, L. Xie, and C. Spanos, "Freedetector: Device-free occupancy detection with commodity wifi," in *IEEE International Conference on Sensing, Communication and Networking (SECON Workshops)*, 2017.
22. J. Yang, **J. Yang**, S. Wang, and H. Zhong, "Learning to Recognize Blurred Objects," *Under Review*.

Awards

1. **Winner**, "DIGIX Global AI Challenge – Digital Device Image Retrieval," *Nanjing, China*, 2020.

2. **3rd Prize Winner**, "National AI Challenge – Person Re-Identification," *Shenzhen, China*, 2020.
3. **Rank 3rd**, "IEEE CVPR 2020 EmotionNet Challenge," *Online*, 2020.
4. **Winner**, "IEEE CVPR 2019 UG2+ Challenge," *Long Beach, USA*, 2019.
5. **Winner**, "HUAWEI Cloud Garbage Recognition Challenge," *Shanghai, China*, 2019.
6. **Winner**, "ACM ICMI 2018 Emotion Recognition Challenge," *Colorado, USA*, 2019.
7. **Winner**, "PEER Hub ImageNet Challenge by Pacific Earthquake Engineering Research Center," *Los Angeles, USA*, 2018.
8. **Rank 2nd**, "Microsoft Imagine Cup (China)," *Shanghai, China*, 2018.
9. **Winner**, "Buildathon of Singapore by Ministry of Education," *Singapore*, 2017.
10. **Winner**, "China Mobile Internet of Things Hackathons," *Guangzhou, China*, 2017.

Patents and Technical Disclosure

- **J. Yang***, L. Xie*, H. Zou, and H. Jiang, "Device-Free Sedentary Period Detection Using Commodity WiFi Signals," *Technical Disclosure, NTU Ref: 2018-127*, April. 2018.
- H. Zou, Y. Zhou, C. Spanos, **J. Yang**, and L. Xie, "Device-Free Human Identification and Device-Free Gesture Recognition," *International Patent, PCT/US19/47049, WO2020037313*, Feb. 2020.
- H. Zou, Y. Zhou, C. Spanos, and **J. Yang**, "Unsupervised WiFi-Enabled Device-User Association for Personalized Location-based Service," *International Patent, PCT/US2019/033782, WO2019226910A1*, Nov. 2019.
- C. Wang*, L. Xie*, **J. Yang**, and J. Yuan, "Kervolutional Neural Networks," *NTU Ref: PAT/344/17/18/SG PRV, Singapore Provisional Patent Application No.: 10201802135Y.*, March. 2018.

Bibliography

- [1] Yann LeCun, Léon Bottou, Yoshua Bengio, and Patrick Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998. [xxi](#), [13](#), [14](#), [91](#)
- [2] Yaroslav Ganin and Victor Lempitsky. Unsupervised domain adaptation by backpropagation. In Francis Bach and David Blei, editors, *Proceedings of the 32nd International Conference on Machine Learning*, volume 37 of *Proceedings of Machine Learning Research*, pages 1180–1189, Lille, France, 07–09 Jul 2015. PMLR. [xxii](#), [20](#), [22](#), [24](#), [91](#), [92](#), [93](#), [99](#), [107](#), [108](#), [110](#), [111](#), [113](#), [116](#), [119](#), [123](#), [124](#), [126](#), [127](#), [128](#), [156](#)
- [3] Mingsheng Long, Yue Cao, Jianmin Wang, and Michael I. Jordan. Learning transferable features with deep adaptation networks. In *Proceedings of the 32Nd International Conference on International Conference on Machine Learning - Volume 37*, ICML’15, pages 97–105, 2015. [xxiv](#), [19](#), [20](#), [25](#), [90](#), [91](#), [92](#), [93](#), [95](#), [108](#), [110](#), [124](#), [127](#), [128](#), [141](#), [146](#), [150](#)
- [4] Mordor Intelligence. *IoT sensor market - growth, trends, and forecast (2020-2025)*, 2019. URL <https://www.mordorintelligence.com/industry-reports/iot-sensor-market>. [1](#)
- [5] Zhenghua Chen, Qingchang Zhu, Yeng Chai Soh, and Le Zhang. Robust human activity recognition using smartphone sensors via ct-pca and online svm. *IEEE Transactions on Industrial Informatics*, 13(6):3070–3080, 2017. [1](#), [30](#)
- [6] Stephan Sigg, Shuyu Shi, and Yusheng Ji. Rf-based device-free recognition of simultaneously conducted activities. In *Proceedings of the 2013 ACM conference on Pervasive and ubiquitous computing adjunct publication*, pages 531–540, 2013. [1](#), [11](#)
- [7] Robert H Dodier, Gregor P Henze, Dale K Tiller, and Xin Guo. Building occupancy detection through sensor belief networks. *Energy and buildings*, 38(9):1033–1043, 2006. [2](#), [10](#)
- [8] S Szewczyk, K Dwan, B Minor, B Swedlove, and D Cook. Annotating smart environment sensor data for activity learning. *Technology and Health Care*, 17(3):161–169, 2009. [2](#)

- [9] Johan AK Suykens and Joos Vandewalle. Least squares support vector machine classifiers. *Neural processing letters*, 9(3):293–300, 1999. [2](#)
- [10] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *European conference on computer vision*, pages 740–755. Springer, 2014. [2](#), [10](#), [13](#)
- [11] Yann LeCun, Yoshua Bengio, and Geoffrey Hinton. Deep learning. *nature*, 521(7553):436–444, 2015. [2](#), [14](#)
- [12] Sinno Jialin Pan, Qiang Yang, et al. A survey on transfer learning. *IEEE Transactions on knowledge and data engineering*, 22(10):1345–1359, 2010. [3](#), [17](#), [22](#), [135](#)
- [13] Gil Friedrich and Yehezkel Yeshurun. Seeing people in the dark: Face recognition in infrared images. In *International Workshop on Biologically Motivated Computer Vision*, pages 348–359. Springer, 2002. [10](#)
- [14] Mahmoud Moghavvemi and Lu Chin Seng. Pyroelectric infrared sensor for intruder detection. In *2004 IEEE Region 10 Conference TENCN 2004.*, volume 500, pages 656–659. IEEE, 2004. [10](#)
- [15] Nishkam Ravi, Nikhil Dandekar, Preetham Mysore, and Michael L Littman. Activity recognition from accelerometer data. In *AAAI*, volume 5, pages 1541–1546, 2005. [10](#), [30](#)
- [16] Jennifer R Kwapisz, Gary M Weiss, and Samuel A Moore. Activity recognition using cell phone accelerometers. *ACM SigKDD Explorations Newsletter*, 12(2):74–82, 2011. [10](#)
- [17] Jindong Wang, Yiqiang Chen, Lisha Hu, Xiaohui Peng, and S Yu Philip. Stratified transfer learning for cross-domain activity recognition. In *2018 IEEE International Conference on Pervasive Computing and Communications (PerCom)*, pages 1–10. IEEE, 2018. [10](#)
- [18] Shuangquan Wang and Gang Zhou. A review on radio based activity recognition. *Digital Communications and Networks*, 1(1):20–29, 2015. [11](#)
- [19] Hongbo Gao, Bo Cheng, Jianqiang Wang, Keqiang Li, Jianhui Zhao, and Deyi Li. Object classification using cnn-based fusion of vision and lidar in autonomous vehicle environment. *IEEE Transactions on Industrial Informatics*, 14(9):4224–4231, 2018. [11](#)
- [20] Bichen Wu, Alvin Wan, Xiangyu Yue, and Kurt Keutzer. Squeezeseg: Convolutional neural nets with recurrent crf for real-time road-object segmentation from 3d lidar point cloud. In *2018 IEEE International Conference on Robotics and Automation (ICRA)*, pages 1887–1893. IEEE, 2018. [11](#)

- [21] Dimitris Zermas, Izzat Izzat, and Nikolaos Papanikolopoulos. Fast segmentation of 3d point clouds: A paradigm on lidar data for autonomous vehicle applications. In *2017 IEEE International Conference on Robotics and Automation (ICRA)*, pages 5067–5073. IEEE, 2017. [11](#)
- [22] Joel D Hewlett, Milos Manic, and Craig G Rieger. Wesbes: A wireless embedded sensor for improving human comfort metrics using temporospatially correlated data. In *2012 5th International Symposium on Resilient Control Systems*, pages 31–36. IEEE, 2012. [11](#)
- [23] Luis M Candanedo and Véronique Feldheim. Accurate occupancy detection of an office room from light, temperature, humidity and co 2 measurements using statistical learning models. *Energy and Buildings*, 112:28–39, 2016. [11](#), [30](#), [37](#)
- [24] Daniel Halperin, Wenjun Hu, Anmol Sheth, and David Wetherall. Tool release: Gathering 802.11 n traces with channel state information. *ACM SIGCOMM Computer Communication Review*, 41(1):53–53, 2011. [12](#), [31](#), [34](#), [35](#), [58](#)
- [25] Yaxiong Xie, Zhenjiang Li, and Mo Li. Precise power delay profiling with commodity wifi. In *Proceedings of the 21st Annual International Conference on Mobile Computing and Networking*, pages 53–64. ACM, 2015. [12](#), [35](#)
- [26] Xuyu Wang, Lingjun Gao, Shiwen Mao, and Santosh Pandey. Csi-based fingerprinting for indoor localization: A deep learning approach. *IEEE Transactions on Vehicular Technology*, 66(1):763–776, 2017. [12](#), [62](#)
- [27] Wei Wang, Alex X Liu, Muhammad Shahzad, Kang Ling, and Sanglu Lu. Device-free human activity recognition using commercial wifi devices. *IEEE Journal on Selected Areas in Communications*, 35(5):1118–1131, 2017. [12](#), [64](#)
- [28] Jianfei Yang, Han Zou, Hao Jiang, and Lihua Xie. Carefi: Sedentary behavior monitoring system via commodity wifi infrastructures. *IEEE Transactions on Vehicular Technology*, 2018. [12](#)
- [29] Han Zou, Yuxun Zhou, Jianfei Yang, Weixi Gu, Lihua Xie, and Costas Spanos. Wifi-based human identification via convex tensor shaplet learning. In *AAAI 2018*, 2018. [12](#), [41](#), [62](#)
- [30] Xiaolong Zheng, Jiliang Wang, Longfei Shangguan, Zimu Zhou, and Yunhao Liu. Smokey: Ubiquitous smoking detection with commercial wifi infrastructures. In *Proceedings of IEEE INFOCOM*, 2016. [12](#), [47](#)
- [31] Wei Xi, Jizhong Zhao, Xiang-Yang Li, Kun Zhao, Shaojie Tang, Xue Liu, and Zhiping Jiang. Electronic frog eye: Counting crowd using wifi. In *Proceedings of IEEE INFOCOM*, 2014. [12](#)

- [32] Han Zou, Yuxun Zhou, Jianfei Yang, Weixi Gu, Lihua Xie, and Costas Spanos. Freedetector: Device-free occupancy detection with commodity wifi. In *Sensing, Communication and Networking (SECON Workshops), 2017 IEEE International Conference on*, pages 1–5. IEEE, 2017. [12](#)
- [33] Yan Wang, Jian Liu, Yingying Chen, Marco Gruteser, Jie Yang, and Hongbo Liu. E-eyes: device-free location-oriented activity identification using fine-grained wifi signatures. In *Proceedings of the 20th annual international conference on Mobile computing and networking*, pages 617–628. ACM, 2014. [12](#), [56](#)
- [34] Kamran Ali, Alex X Liu, Wei Wang, and Muhammad Shahzad. Keystroke recognition using wifi signals. In *Proceedings of ACM MobiCom*, 2015. [12](#), [44](#)
- [35] Han Zou, Yuxun Zhou, Jianfei Yang, Hao Jiang, Lihua Xie, and Costas J Spanos. Deepsense: Device-free human activity recognition via autoencoder long-term recurrent convolutional network. In *2018 IEEE International Conference on Communications (ICC)*, pages 1–6. IEEE, 2018. [12](#)
- [36] Jie Zhang, Zhanyong Tang, Meng Li, Dingyi Fang, Petteri Nurmi, and Zheng Wang. Crosssense: Towards cross-site and large-scale wifi sensing. In *Proceedings of the 24th Annual International Conference on Mobile Computing and Networking, MobiCom '18*, pages 305–320, New York, NY, USA, 2018. ACM. ISBN 978-1-4503-5903-0. doi: 10.1145/3241539.3241570. URL <http://doi.acm.org/10.1145/3241539.3241570>.
- [37] Wenjun Jiang, Chenglin Miao, Fenglong Ma, Shuochao Yao, Yaqing Wang, Ye Yuan, Hongfei Xue, Chen Song, Xin Ma, Dimitrios Koutsonikolas, et al. Towards environment independent device free human activity recognition. In *Proceedings of the 24th Annual International Conference on Mobile Computing and Networking*, pages 289–304. ACM, 2018.
- [38] Han Zou, Jianfei Yang, Yuxun Zhou, Lihua Xie, and Costas J Spanos. Robust wifi-enabled device-free gesture recognition via unsupervised adversarial domain adaptation. In *2018 27th International Conference on Computer Communication and Networks (ICCCN)*, pages 1–8. IEEE, 2018. [12](#)
- [39] Svante Wold, Kim Esbensen, and Paul Geladi. Principal component analysis. *Chemometrics and intelligent laboratory systems*, 2(1-3):37–52, 1987. [13](#)
- [40] Vladimir N Vapnik. An overview of statistical learning theory. *IEEE transactions on neural networks*, 10(5):988–999, 1999. [13](#)
- [41] Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural computation*, 9(8):1735–1780, 1997. [14](#)
- [42] Jason PC Chiu and Eric Nichols. Named entity recognition with bidirectional lstm-cnns. *Transactions of the Association for Computational Linguistics*, 4: 357–370, 2016. [14](#)

- [43] Jun Liu, Amir Shahroudy, Dong Xu, and Gang Wang. Spatio-temporal lstm with trust gates for 3d human action recognition. In *European conference on computer vision*, pages 816–833. Springer, 2016. [14](#)
- [44] Jake Snell, Kevin Swersky, and Richard Zemel. Prototypical networks for few-shot learning. In *Advances in Neural Information Processing Systems*, pages 4077–4087, 2017. [15](#), [139](#), [141](#)
- [45] Jianfei Yang, Han Zou, Yuxun Zhou, and Lihua Xie. Learning gestures from wifi: A siamese recurrent convolutional architecture. *IEEE Internet of Things Journal*, 6(6):10763–10772, 2019. [15](#)
- [46] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009. [15](#), [125](#)
- [47] Yue Zheng, Yi Zhang, Kun Qian, Guidong Zhang, Yunhao Liu, Chenshu Wu, and Zheng Yang. Zero-effort cross-domain gesture recognition with wi-fi. In *Proceedings of the 17th Annual International Conference on Mobile Systems, Applications, and Services*, pages 313–325, 2019. [15](#)
- [48] Shai Ben-David, John Blitzer, Koby Crammer, Alex Kulesza, Fernando Pereira, and Jennifer Wortman Vaughan. A theory of learning from different domains. *Machine learning*, 79(1-2):151–175, 2010. [15](#), [18](#), [22](#), [97](#), [100](#), [101](#), [106](#), [118](#), [131](#), [143](#), [144](#)
- [49] Sinno Jialin Pan, Vincent Wenchen Zheng, Qiang Yang, and Derek Hao Hu. Transfer learning for wifi-based indoor localization. In *Association for the advancement of artificial intelligence (AAAI) workshop*, page 6, 2008. [16](#), [17](#)
- [50] Mingmin Zhao, Tianhong Li, Mohammad Abu Alsheikh, Yonglong Tian, Hang Zhao, Antonio Torralba, and Dina Katabi. Through-wall human pose estimation using radio signals. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 7356–7365, 2018. [16](#)
- [51] Han Zou, Jianfei Yang, Hari Prasanna Das, Huihan Liu, Yuxun Zhou, and Costas J Spanos. Wifi and vision multimodal learning for accurate and robust device-free human activity recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pages 0–0, 2019. [16](#)
- [52] Shangyu Chen, Wenya Wang, and Sinno Jialin Pan. Cooperative pruning in cross-domain deep neural network compression. In *Proceedings of the 28th International Joint Conference on Artificial Intelligence*, pages 2102–2108. AAAI Press, 2019. [17](#), [84](#)
- [53] Jianfei Yang, Han Zou, Shuxin Cao, Zhenghua Chen, and Lihua Xie. Mobileda: Towards edge domain adaptation. *IEEE Internet of Things Journal*, 2020. [17](#)

- [54] Han Zou, Yuxun Zhou, Hao Jiang, Baoqi Huang, Lihua Xie, and Costas Spanos. Adaptive localization in dynamic indoor environments by transfer kernel learning. In *Wireless Communications and Networking Conference (WCNC), 2017 IEEE*, pages 1–6. IEEE, 2017. [17](#)
- [55] Arthur Gretton, Karsten M Borgwardt, Malte Rasch, Bernhard Schölkopf, and Alex J Smola. A kernel method for the two-sample-problem. In *Advances in neural information processing systems*, pages 513–520, 2007. [18](#), [83](#), [99](#)
- [56] Sinno Jialin Pan, Ivor W Tsang, James T Kwok, and Qiang Yang. Domain adaptation via transfer component analysis. *IEEE Transactions on Neural Networks*, 22(2):199–210, 2010. [19](#), [20](#)
- [57] Arthur Gretton, Dino Sejdinovic, Heiko Strathmann, Sivaraman Balakrishnan, Massimiliano Pontil, Kenji Fukumizu, and Bharath K Sriperumbudur. Optimal kernel choice for large-scale two-sample tests. In *NIPS*, pages 1205–1213, 2012. [19](#), [68](#), [69](#), [87](#), [91](#), [96](#), [99](#)
- [58] Eric Tzeng, Judy Hoffman, Ning Zhang, Kate Saenko, and Trevor Darrell. Deep domain confusion: Maximizing for domain invariance. *CoRR*, abs/1412.3474, 2014. URL <http://arxiv.org/abs/1412.3474>. [19](#), [22](#), [116](#), [119](#), [124](#), [126](#), [128](#), [146](#)
- [59] Mingsheng Long, Han Zhu, Jianmin Wang, and Michael I Jordan. Deep transfer learning with joint adaptation networks. In *Proceedings of the 34th International Conference on Machine Learning-Volume 70*, pages 2208–2217. JMLR. org, 2017. [19](#)
- [60] Kenji Fukumizu, Arthur Gretton, Gert R Lanckriet, Bernhard Schölkopf, and Bharath K Sriperumbudur. Kernel choice and classifiability for rkhs embeddings of probability distributions. In *Advances in neural information processing systems*, pages 1750–1758, 2009. [19](#)
- [61] Baochen Sun and Kate Saenko. Deep coral: Correlation alignment for deep domain adaptation. In *European Conference on Computer Vision*, pages 443–450. Springer, 2016. [20](#), [81](#), [82](#), [83](#), [87](#), [90](#), [92](#), [95](#), [96](#), [108](#), [110](#), [116](#), [124](#), [126](#), [146](#)
- [62] Mahsa Baktashmotlagh, Mehrtash T Harandi, Brian C Lovell, and Mathieu Salzmann. Domain adaptation on the statistical manifold. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2481–2488, 2014. [20](#)
- [63] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In *Advances in neural information processing systems*, pages 2672–2680, 2014. [20](#), [83](#), [116](#)

- [64] Yuchen Zhang, Tianle Liu, Mingsheng Long, and Michael Jordan. Bridging theory and algorithm for domain adaptation. In *International Conference on Machine Learning*, pages 7404–7413, 2019. [20](#)
- [65] Zirui Wang, Zihang Dai, Barnabas Poczos, and Jaime Carbonell. Characterizing and avoiding negative transfer. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2019. [20](#)
- [66] Judy Hoffman, Eric Tzeng, Taesung Park, Jun-Yan Zhu, Phillip Isola, Kate Saenko, Alexei Efros, and Trevor Darrell. CyCADA: Cycle-consistent adversarial domain adaptation. In *Proceedings of the 35th International Conference on Machine Learning*, pages 1989–1998, 2018. [20](#), [22](#), [100](#), [108](#), [110](#), [124](#), [126](#)
- [67] Konstantinos Bousmalis, Nathan Silberman, David Dohan, Dumitru Erhan, and Dilip Krishnan. Unsupervised pixel-level domain adaptation with generative adversarial networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3722–3731, 2017. [20](#), [22](#)
- [68] John Blitzer, Koby Crammer, Alex Kulesza, Fernando Pereira, and Jennifer Wortman. Learning bounds for domain adaptation. In *Advances in neural information processing systems*, pages 129–136, 2008. [21](#)
- [69] Sinno Jialin Pan, Ivor W Tsang, James T Kwok, and Qiang Yang. Domain adaptation via transfer component analysis. *Neural Networks, IEEE Transactions on*, 22(2):199–210, 2011. [22](#), [23](#), [90](#), [92](#), [116](#)
- [70] Eric Tzeng, Judy Hoffman, Kate Saenko, and Trevor Darrell. Adversarial discriminative domain adaptation. *arXiv preprint arXiv:1702.05464*, 2017. [22](#), [83](#), [90](#), [91](#), [92](#), [107](#), [108](#), [110](#), [124](#), [125](#), [126](#)
- [71] Xinyang Chen, Sinan Wang, Mingsheng Long, and Jianmin Wang. Transferability vs. discriminability: Batch spectral penalization for adversarial domain adaptation. In Kamalika Chaudhuri and Ruslan Salakhutdinov, editors, *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pages 1081–1090, Long Beach, California, USA, 09–15 Jun 2019. PMLR. [22](#), [116](#), [119](#), [120](#), [124](#), [126](#), [127](#), [128](#)
- [72] Rui Shu, Hung H Bui, Hirokazu Narui, and Stefano Ermon. A dirt-t approach to unsupervised domain adaptation. In *Proc. 6th International Conference on Learning Representations*, 2018. [22](#), [24](#), [100](#), [116](#)
- [73] Kuniaki Saito, Kohei Watanabe, Yoshitaka Ushiku, and Tatsuya Harada. Maximum classifier discrepancy for unsupervised domain adaptation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3723–3732, 2018. [22](#), [24](#), [108](#), [110](#), [116](#), [124](#), [126](#), [131](#), [146](#), [147](#)
- [74] Shaoan Xie, Zibin Zheng, Liang Chen, and Chuan Chen. Learning semantic representations for unsupervised domain adaptation. In *Proceedings of*

- the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pages 5423–5432, Stockholmsmssan, Stockholm Sweden, 10–15 Jul 2018. PMLR. [22](#), [116](#)
- [75] Yaroslav Ganin and Victor Lempitsky. Unsupervised domain adaptation by backpropagation. In *ICML*, pages 1180–1189, 2015. [23](#), [146](#)
- [76] Kuniaki Saito, Yoshitaka Ushiku, and Tatsuya Harada. Asymmetric tri-training for unsupervised domain adaptation. In *Proceedings of the 34th International Conference on Machine Learning-Volume 70*, pages 2988–2997. JMLR. org, 2017. [23](#)
- [77] Shaoan Xie, Zibin Zheng, Liang Chen, and Chuan Chen. Learning semantic representations for unsupervised domain adaptation. In *International Conference on Machine Learning*, pages 5419–5428, 2018. [141](#), [146](#)
- [78] Yingwei Pan, Ting Yao, Yehao Li, Yu Wang, Chong-Wah Ngo, and Tao Mei. Transferrable prototypical networks for unsupervised domain adaptation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2239–2247, 2019. [23](#), [143](#), [146](#)
- [79] Pedro O Pinheiro. Unsupervised domain adaptation with similarity learning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 8004–8013, 2018. [23](#)
- [80] Yen-Chang Hsu, Zhaoyang Lv, and Zsolt Kira. Learning to cluster in order to transfer across domains and tasks. In *International Conference on Learning Representations (ICLR)*, 2018. [23](#), [124](#), [125](#), [127](#)
- [81] Ziwei Liu, Zhongqi Miao, Xiaohang Zhan, Jiayun Wang, Boqing Gong, and Stella X Yu. Large-scale long-tailed recognition in an open world. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2537–2546, 2019. [23](#), [136](#)
- [82] Xiao Zhang, Zhiyuan Fang, Yandong Wen, Zhifeng Li, and Yu Qiao. Range loss for deep face recognition with long-tailed training data. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 5409–5418, 2017. [23](#)
- [83] Zachary C Lipton, Yu-Xiang Wang, and Alex Smola. Detecting and correcting for label shift with black box predictors. *arXiv preprint arXiv:1802.03916*, 2018. [23](#)
- [84] Kamyar Azizzadenesheli, Anqi Liu, Fanny Yang, and Animashree Anandkumar. Regularized learning for domain adaptation under label shifts. *arXiv preprint arXiv:1903.09734*, 2019. [23](#)
- [85] Hongliang Yan, Yukang Ding, Peihua Li, Qilong Wang, Yong Xu, and Wangmeng Zuo. Mind the class weight bias: Weighted maximum mean discrepancy for unsupervised domain adaptation. In *Proceedings of the IEEE Conference*

- on Computer Vision and Pattern Recognition*, pages 2272–2281, 2017. [23](#), [146](#)
- [86] Pau Panareda Busto and Juergen Gall. Open set domain adaptation. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 754–763, 2017. [23](#)
- [87] Zhangjie Cao, Lijia Ma, Mingsheng Long, and Jianmin Wang. Partial adversarial domain adaptation. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 135–150, 2018. [23](#), [146](#), [147](#)
- [88] Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A Efros. Unpaired image-to-image translation using cycle-consistent adversarial networks. In *Proceedings of the IEEE international conference on computer vision*, pages 2223–2232, 2017. [23](#), [107](#)
- [89] Muhammad Ghifary, W Bastiaan Kleijn, Mengjie Zhang, David Balduzzi, and Wen Li. Deep reconstruction-classification networks for unsupervised domain adaptation. In *European Conference on Computer Vision*, pages 597–613. Springer, 2016. [23](#), [90](#), [91](#), [108](#), [110](#), [124](#), [126](#)
- [90] Konstantinos Bousmalis, George Trigeorgis, Nathan Silberman, Dilip Krishnan, and Dumitru Erhan. Domain separation networks. In *Advances in Neural Information Processing Systems*, pages 343–351, 2016. [23](#), [107](#), [108](#), [110](#), [124](#), [126](#)
- [91] Xavier Glorot, Antoine Bordes, and Yoshua Bengio. Domain adaptation for large-scale sentiment classification: A deep learning approach. In *Proceedings of the 28th international conference on machine learning (ICML-11)*, pages 513–520, 2011. [23](#)
- [92] Abhishek Kumar, Prasanna Sattigeri, Kahini Wadhawan, Leonid Karlinsky, Rogerio Feris, Bill Freeman, and Gregory Wornell. Co-regularized alignment for unsupervised domain adaptation. In *Advances in Neural Information Processing Systems*, pages 9345–9356, 2018. [24](#)
- [93] Yaroslav Ganin, Evgeniya Ustinova, Hana Ajakan, Pascal Germain, Hugo Larochelle, François Laviolette, Mario Marchand, and Victor Lempitsky. Domain-adversarial training of neural networks. *The Journal of Machine Learning Research*, 17(1):2096–2030, 2016. [24](#), [90](#), [96](#), [97](#), [102](#), [110](#), [124](#)
- [94] Kate Saenko, Brian Kulis, Mario Fritz, and Trevor Darrell. Adapting visual category models to new domains. In *European conference on computer vision*, pages 213–226. Springer, 2010. [24](#), [89](#), [145](#)
- [95] Hemanth Venkateswara, Jose Eusebio, Shayok Chakraborty, and Sethuraman Panchanathan. Deep hashing network for unsupervised domain adaptation. In *Proc. CVPR*, pages 5018–5027, 2017. [25](#), [145](#)

- [96] Xingchao Peng, Ben Usman, Neela Kaushik, Dequan Wang, Judy Hoffman, and Kate Saenko. Visda: A synthetic-to-real benchmark for visual domain adaptation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pages 2021–2026, 2018. [25](#)
- [97] Han Zou, Yuxun Zhou, Jianfei Yang, Huihan Liu, Hari Prasanna Das, and Costas J Spanos. Consensus adversarial domain adaptation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 5997–6004, 2019. [25](#), [91](#), [108](#), [110](#), [116](#)
- [98] Yaxiong Xie, Zhenjiang Li, and Mo Li. Precise power delay profiling with commodity wi-fi. *IEEE Transactions on Mobile Computing*, 18(6):1342–1355, 2018. [25](#)
- [99] Junjing Yang, Mattheos Santamouris, and Siew Eang Lee. Review of occupancy sensing systems and occupancy modeling methodologies for the application in institutional buildings. *Energy and Buildings*, 121:344–349, 2016. [30](#)
- [100] Lars Zimmermann, Robert Weigel, and Georg Fischer. Fusion of non-intrusive environmental sensors for occupancy detection in smart homes. *IEEE Internet of Things Journal*, pages 1–1, 2017. doi: 10.1109/JIOT.2017.2752134. [30](#)
- [101] Limin Wang, Yu Qiao, and Xiaoou Tang. Action recognition with trajectory-pooled deep-convolutional descriptors. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4305–4314, 2015. [30](#)
- [102] Mingmin Zhao, Shichao Yue, Dina Katabi, Tommi S Jaakkola, and Matt T Bianchi. Learning sleep stages from radio signals: A conditional adversarial architecture. In *International Conference on Machine Learning*, pages 4100–4109, 2017. [30](#)
- [103] Yu Gu, Fuji Ren, and Jie Li. Paws: Passive human activity recognition based on wifi ambient signals. *IEEE Internet of Things Journal*, 3(5):796–805, 2016. [31](#)
- [104] Han Zou, Yuxun Zhou, Jianfei Yang, Weixi Gu, Lihua Xie, and Costas Spanos. Poster: Wifi-based device-free human activity recognition via automatic representation learning. In *Proceedings of the 23rd Annual International Conference on Mobile Computing and Networking*, pages 606–608. ACM, 2017. [31](#)
- [105] Han Zou, Yuxun Zhou, Jianfei Yang, Weixi Gu, Lihua Xie, and Costas Spanos. Multiple kernel representation learning for wifi-based human activity recognition. In *Machine Learning and Applications (ICMLA), 2017 16th IEEE International Conference on*, pages 268–274. IEEE, 2017. [31](#)

- [106] Zheng Yang, Zimu Zhou, and Yunhao Liu. From rssi to csi: Indoor localization via channel response. *ACM Computing Surveys (CSUR)*, 46(2):25, 2013. [32](#)
- [107] Jon Gjengset, Jie Xiong, Graeme McPhillips, and Kyle Jamieson. Phaser: Enabling phased array signal processing on commodity wifi access points. In *Proceedings of the 20th annual international conference on Mobile computing and networking*, pages 153–164. ACM, 2014. [33](#)
- [108] Souvik Sen, Jeongkeun Lee, Kyu-Han Kim, and Paul Congdon. Avoiding multipath to revive inbuilding wifi localization. In *Proceeding of the 11th annual international conference on Mobile systems, applications, and services*, pages 249–262. ACM, 2013. [34](#)
- [109] Ju Wang, Hongbo Jiang, Jie Xiong, Kyle Jamieson, Xiaojiang Chen, Dingyi Fang, and Binbin Xie. Lif: low human-effort, device-free localization with fine-grained subcarrier information. In *Proceedings of the 22nd Annual International Conference on Mobile Computing and Networking*, pages 243–256. ACM, 2016. [35](#)
- [110] Urs Hunkeler, Hong Linh Truong, and Andy Stanford-Clark. Mqtt-sa publish/subscribe protocol for wireless sensor networks. In *Communication systems software and middleware and workshops, 2008. comsware 2008. 3rd international conference on*, pages 791–798. IEEE, 2008. [36](#)
- [111] Varick L Erickson, Miguel Á Carreira-Perpiñán, and Alberto E Cerpa. Occupancy modeling and prediction for building energy management. *ACM Transactions on Sensor Networks (TOSN)*, 10(3):42, 2014. [37](#)
- [112] Han Zou, Baoqi Huang, Xiaoxuan Lu, Hao Jiang, and Lihua Xie. A robust indoor positioning system based on the procrustes analysis and weighted extreme learning machine. *IEEE Transactions on Wireless Communications*, 15(2):1252–1266, 2016. [39](#)
- [113] Ann M Swartz, Aubrianne E Rote, Young Ik Cho, Whitney A Welch, and Scott J Strath. Responsiveness of motion sensors to detect change in sedentary and physical activity behaviour. *Br J Sports Med*, 2014. [40](#), [41](#)
- [114] Stuart JH Biddle, Trish Gorely, Simon J Marshall, Ian Murdey, and Noel Cameron. Physical activity and sedentary behaviours in youth: issues and controversies. *The journal of the Royal Society for the Promotion of Health*, 2004. [40](#)
- [115] Alicia A Thorp, Neville Owen, Maike Neuhaus, and David W Dunstan. Sedentary behaviors and subsequent health outcomes in adults: a systematic review of longitudinal studies, 1996–2011. *American journal of preventive medicine*, 2011. [40](#)

- [116] Sangwon Bae, Anind K Dey, and Carissa A Low. Using passively collected sedentary behavior to predict hospital readmission. In *Proceedings of ACM Ubicomp Adjunct*, 2016. [40](#)
- [117] Jonathan Synnott, Joseph Rafferty, and Chris D Nugent. Detection of workplace sedentary behavior using thermal sensors. In *Proceedings of IEEE International Conference of the Engineering in Medicine and Biology Society (EMBC)*, 2016. [41](#)
- [118] Jianxin Wu, Adebola Osuntogun, Tanzeem Choudhury, Matthai Philipose, and James M Rehg. A scalable approach to activity recognition based on object use. In *ICCV*. IEEE, 2007. [41](#)
- [119] Han Zou, Yuxun Zhou, Jianfei Yang, Weixi Gu, Lihua Xie, and Costas Spanos. Freecount: Device-free crowd counting with commodity wifi. In *2017 IEEE Global Communications Conference (GLOBECOM)*. IEEE, 2017. [42](#)
- [120] Ahmed Elgammal, Ramani Duraiswami, David Harwood, and Larry S Davis. Background and foreground modeling using nonparametric kernel density estimation for visual surveillance. *Proceedings of the IEEE*, 90(7):1151–1163, 2002. [48](#)
- [121] Han Zou, Hao Jiang, Jianfei Yang, Lihua Xie, and Costas Spanos. Non-intrusive occupancy sensing in commercial buildings. *Energy and Buildings*, 154:633–643, 2017. [53](#)
- [122] Han Zou, Yuxun Zhou, Hao Jiang, Szu-Cheng Chien, Lihua Xie, and Costas J Spanos. Winlight: A wifi-based occupancy-driven lighting control system for smart building. *Energy and Buildings*, 158:924–938, 2018. [53](#)
- [123] Wei Wang, Alex X Liu, Muhammad Shahzad, Kang Ling, and Sanglu Lu. Understanding and modeling of wifi signal based human activity recognition. In *Proceedings of the 21st Annual International Conference on Mobile Computing and Networking*, pages 65–76. ACM, 2015. [56](#)
- [124] Xuyu Wang, Xiangyu Wang, and Shiwen Mao. Rf sensing in the internet of things: A general deep learning framework. *IEEE Communications Magazine*, 56(9):62–67, 2018. [62](#)
- [125] Fangxin Wang, Wei Gong, and Jiangchuan Liu. On spatial diversity in wifi-based human activity recognition: A deep learning based approach. *IEEE Internet of Things Journal*, 2018.
- [126] Jianfei Yang, Han Zou, Hao Jiang, and Lihua Xie. Device-free occupant activity sensing using wifi-enabled iot devices for smart homes. *IEEE Internet of Things Journal*, 5(5):3991–4002, 2018. [62](#)
- [127] Aditya Virmani and Muhammad Shahzad. Position and orientation agnostic gesture recognition using wifi. In *Mobisys 2017*, pages 252–264. ACM, 2017. [64](#)

- [128] Gregory Koch, Richard Zemel, and Ruslan Salakhutdinov. Siamese neural networks for one-shot image recognition. In *ICML Deep Learning Workshop*, volume 2, 2015. 65, 68
- [129] Vinod Nair and Geoffrey E Hinton. Rectified linear units improve restricted boltzmann machines. In *Proceedings of the 27th international conference on machine learning (ICML-10)*, pages 807–814, 2010. 66
- [130] Andrew L Maas, Awni Y Hannun, and Andrew Y Ng. Rectifier nonlinearities improve neural network acoustic models. In *ICML*, volume 30, 2013. 66
- [131] Ya Li, Xinmei Tian, Xu Shen, and Dacheng Tao. Classification and representation joint learning via deep networks. In *IJCAI 2017*, 2017. 67
- [132] Jasper Snoek, Hugo Larochelle, and Ryan P Adams. Practical bayesian optimization of machine learning algorithms. In *Advances in neural information processing systems*, pages 2951–2959, 2012. 73
- [133] Wenfeng He, Kaishun Wu, Yongpan Zou, and Zhong Ming. Wig: Wifi-based gesture recognition system. In *Computer Communication and Networks (ICCCN), 2015 24th International Conference on*, pages 1–7. IEEE, 2015. 73
- [134] Mohammed Abdulaziz Aide Al-qaness and Fangmin Li. Wiger: Wifi-based gesture recognition system. *ISPRS International Journal of Geo-Information*, 5(6):92, 2016. 73
- [135] Chen Wang, Le Zhang, Lihua Xie, and Junsong Yuan. Kernel Cross-Correlator. In *AAAI Conference on Artificial Intelligence (AAAI-18)*, pages 4179–4186, 2018. 73
- [136] Ross Girshick, Jeff Donahue, Trevor Darrell, and Jitendra Malik. Rich feature hierarchies for accurate object detection and semantic segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 580–587, 2014. 82
- [137] Jimmy Ba and Rich Caruana. Do deep nets really need to be deep? In *Advances in neural information processing systems*, pages 2654–2662, 2014. 82
- [138] Geoffrey Hinton, Oriol Vinyals, and Jeffrey Dean. Distilling the knowledge in a neural network. In *NIPS Deep Learning and Representation Learning Workshop*, 2015. URL <http://arxiv.org/abs/1503.02531>. 82, 83, 84, 86
- [139] Shangyu Chen, Wenya Wang, and Sinno Jialin Pan. Metaquant: Learning to quantize by learning to penetrate non-differentiable quantization. In *Advances in Neural Information Processing Systems*, pages 3918–3928, 2019. 84
- [140] Mingsheng Long, Han Zhu, Jianmin Wang, and Michael I Jordan. Deep transfer learning with joint adaptation networks. In *International Conference on Machine Learning (ICML)*, 2017. 84, 90, 92, 93, 124, 127

- [141] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016. [84](#), [92](#), [93](#), [110](#), [127](#), [128](#), [146](#)
- [142] Eric Tzeng, Judy Hoffman, Trevor Darrell, and Kate Saenko. Simultaneous deep transfer across domains and tasks. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 4068–4076, 2015. [87](#)
- [143] Dustin Franklin. Nvidia jetson tx2 delivers twice the intelligence to the edge. *NVIDIA Accelerated Computing— Parallel Forall*, 2017. [89](#)
- [144] Boqing Gong, Yuan Shi, Fei Sha, and Kristen Grauman. Geodesic flow kernel for unsupervised domain adaptation. In *Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on*, pages 2066–2073. IEEE, 2012. [90](#), [92](#)
- [145] Mingsheng Long, Zhangjie Cao, Jianmin Wang, and Michael I Jordan. Conditional adversarial domain adaptation. In *Advances in neural information processing systems*, 2018. [90](#), [92](#), [93](#), [100](#), [108](#), [110](#), [124](#), [126](#), [127](#), [146](#)
- [146] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems*, pages 1097–1105, 2012. [91](#), [92](#), [93](#), [146](#)
- [147] Laurens van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. *Journal of machine learning research*, 9(Nov):2579–2605, 2008. [97](#), [131](#), [150](#)
- [148] Xinyang Chen, Sinan Wang, Mingsheng Long, and Jianmin Wang. Transferability vs. discriminability: Batch spectral penalization for adversarial domain adaptation. In *International Conference on Machine Learning*, pages 1081–1090, 2019. [100](#), [108](#), [110](#)
- [149] Hong Liu, Mingsheng Long, Jianmin Wang, and Michael Jordan. Transferable adversarial training: A general approach to adapting deep classifiers. In *International Conference on Machine Learning*, pages 4013–4022, 2019. [100](#), [120](#)
- [150] Nello Cristianini, John Shawe-Taylor, et al. *An introduction to support vector machines and other kernel-based learning methods*. Cambridge university press, 2000. [102](#)
- [151] Hidetoshi Shimodaira. Improving predictive inference under covariate shift by weighting the log-likelihood function. *Journal of statistical planning and inference*, 90(2):227–244, 2000. [103](#)
- [152] Junbo Zhao, Michael Mathieu, and Yann LeCun. Energy-based generative adversarial network. In *Proc. 5th International Conference on Learning Representations*, 2017. [105](#), [123](#)

- [153] Ximei Wang, Ying Jin, Mingsheng Long, Jianmin Wang, and Michael I Jordan. Transferable normalization: Towards improving transferability of deep neural networks. In *Advances in Neural Information Processing Systems*, pages 1951–1961, 2019. [108](#), [110](#)
- [154] Gustavo Malkomes, Charles Schaff, and Roman Garnett. Bayesian optimization for automated model selection. In *Advances in Neural Information Processing Systems*, pages 2900–2908, 2016. [108](#)
- [155] Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, et al. Scikit-learn: Machine learning in python. *Journal of machine learning research*, 12(Oct):2825–2830, 2011. [111](#)
- [156] Tuan-Hung Vu, Himalaya Jain, Maxime Bucher, Matthieu Cord, and Patrick Perez. Advent: Adversarial entropy minimization for domain adaptation in semantic segmentation. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2019. [116](#)
- [157] Zheng Li, Yun Zhang, Ying Wei, Yuxiang Wu, and Qiang Yang. End-to-end adversarial memory network for cross-domain sentiment classification. In *IJCAI*, pages 2237–2243, 2017. [116](#)
- [158] Yann LeCun, Sumit Chopra, Raia Hadsell, M Ranzato, and F Huang. A tutorial on energy-based learning. *Predicting structured data*, 1(0), 2006. [117](#), [123](#)
- [159] Ming-Yu Liu and Onel Tuzel. Coupled generative adversarial networks. In *Advances in neural information processing systems*, pages 469–477, 2016. [124](#), [126](#)
- [160] Swami Sankaranarayanan, Yogesh Balaji, Carlos D Castillo, and Rama Chellappa. Generate to adapt: Aligning domains using generative adversarial networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 8503–8512, 2018. [124](#), [126](#)
- [161] Jian Liang, Ran He, Zhenan Sun, and Tieniu Tan. Distant supervised centroid shift: A simple and efficient approach to visual domain adaptation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2975–2984, 2019. [124](#), [126](#)
- [162] William J Reed. The pareto, zipf and other power laws. *Economics letters*, 74(1):15–19, 2001. [136](#), [144](#)
- [163] B Schölkopf, D Janzing, J Peters, E Sgouritsa, K Zhang, and J Mooij. On causal and anticausal learning. In *29th International Conference on Machine Learning (ICML 2012)*, pages 1255–1262. International Machine Learning Society, 2012. [136](#)

- [164] Han Zhao, Remi Tachet Des Combes, Kun Zhang, and Geoffrey Gordon. On learning invariant representations for domain adaptation. In *International Conference on Machine Learning*, pages 7523–7532, 2019. [136](#), [138](#), [148](#)
- [165] Zirui Wang, Zihang Dai, Barnabás Póczos, and Jaime Carbonell. Characterizing and avoiding negative transfer. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 11293–11302, 2019. [136](#)
- [166] Olivier Chapelle and Alexander Zien. Semi-supervised classification by low density separation. In *AISTATS*, volume 2005, pages 57–64. Citeseer, 2005. [138](#), [142](#)
- [167] Yen-Chang Hsu and Zsolt Kira. Neural network-based clustering using pairwise constraints. In *ICLR workshop*, 2016. URL <https://arxiv.org/abs/1511.06321>. [139](#)
- [168] Zhijie Deng, Yucen Luo, and Jun Zhu. Cluster alignment with a teacher for unsupervised domain adaptation. In *The IEEE International Conference on Computer Vision (ICCV)*, October 2019. [141](#), [146](#)
- [169] Yves Grandvalet and Yoshua Bengio. Semi-supervised learning by entropy minimization. In *Advances in neural information processing systems*, pages 529–536, 2005. [142](#)
- [170] Mingsheng Long, Yue Cao, Jianmin Wang, and Michael Jordan. Learning transferable features with deep adaptation networks. In *International Conference on Machine Learning*, pages 97–105, 2015. [143](#)
- [171] Shai Ben-David, John Blitzer, Koby Crammer, and Fernando Pereira. Analysis of representations for domain adaptation. In *Advances in neural information processing systems*, pages 137–144, 2007. [144](#)
- [172] Yuval Netzer, Tao Wang, Adam Coates, Alessandro Bissacco, Bo Wu, and Andrew Y Ng. Reading digits in natural images with unsupervised feature learning. In *NIPS workshop on deep learning and unsupervised feature learning*, volume 2011, page 5, 2011. [145](#)
- [173] Flood Sung, Yongxin Yang, Li Zhang, Tao Xiang, Philip HS Torr, and Timothy M Hospedales. Learning to compare: Relation network for few-shot learning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1199–1208, 2018. [150](#)
- [174] Hao Wang, Daqing Zhang, Yasha Wang, Junyi Ma, Yuxiang Wang, and Shengjie Li. Rt-fall: A real-time and contactless fall detection system with commodity wifi devices. *IEEE Transactions on Mobile Computing*, 16(2): 511–526, 2016. [155](#)
- [175] Matthias Schulz, Jakob Link, Francesco Gringoli, and Matthias Hollick. Shadow wi-fi: Teaching smartphones to transmit raw signals and to extract

- channel state information to implement practical covert channels over wi-fi. In *Proceedings of the 16th Annual International Conference on Mobile Systems, Applications, and Services*, pages 256–268, 2018. [155](#)
- [176] Sinno Jialin Pan, Xiaochuan Ni, Jian-Tao Sun, Qiang Yang, and Zheng Chen. Cross-domain sentiment classification via spectral feature alignment. In *Proceedings of the 19th international conference on World wide web*, pages 751–760, 2010. [156](#)
- [177] Wei Zhang, Gaoliang Peng, Chuanhao Li, Yuanhang Chen, and Zhujun Zhang. A new deep learning model for fault diagnosis with good anti-noise and domain adaptation ability on raw vibration signals. *Sensors*, 17(2):425, 2017. [156](#)
- [178] Chelsea Finn, Pieter Abbeel, and Sergey Levine. Model-agnostic meta-learning for fast adaptation of deep networks. In *Proceedings of the 34th International Conference on Machine Learning-Volume 70*, pages 1126–1135. JMLR. org, 2017. [156](#)
- [179] Supriyo Chakraborty, Richard Tomsett, Ramya Raghavendra, Daniel Harborne, Moustafa Alzantot, Federico Cerutti, Mani Srivastava, Alun Preece, Simon Julier, Raghuveer M Rao, et al. Interpretability of deep learning models: a survey of results. In *2017 IEEE SmartWorld, Ubiquitous Intelligence & Computing, Advanced & Trusted Computed, Scalable Computing & Communications, Cloud & Big Data Computing, Internet of People and Smart City Innovation (SmartWorld/SCALCOM/UIC/ATC/CBDCom/IOP/SCI)*, pages 1–6. IEEE, 2017. [156](#)
- [180] Luigi Atzori, Antonio Iera, and Giacomo Morabito. The internet of things: A survey. *Computer networks*, 54(15):2787–2805, 2010. [157](#)
- [181] Amanda Purington, Jessie G Taft, Shruti Sannon, Natalya N Bazarova, and Samuel Hardman Taylor. ” alexa is my new bff” social roles, user satisfaction, and personification of the amazon echo. In *Proceedings of the 2017 CHI Conference Extended Abstracts on Human Factors in Computing Systems*, pages 2853–2859, 2017. [157](#)
- [182] Ilche Georgievski, Tuan Anh Nguyen, and Marco Aiello. Combining activity recognition and ai planning for energy-saving offices. In *2013 IEEE 10th International Conference on Ubiquitous Intelligence and Computing and 2013 IEEE 10th International Conference on Autonomic and Trusted Computing*, pages 238–245. IEEE, 2013. [157](#)
- [183] Tadas Baltrušaitis, Chaitanya Ahuja, and Louis-Philippe Morency. Multi-modal machine learning: A survey and taxonomy. *IEEE transactions on pattern analysis and machine intelligence*, 41(2):423–443, 2018. [157](#)