Visual Cross-Domain Adaptation Under Various Data Access Privileges

an abstract
submitted on the twenty-sixth day of April, 2023
to the Department of Computer Science
of the School of Science and Engineering of
Tulane University
in partial fulfillment of the requirements
for the degree of
Doctor of Philosophy
by

Haifeng Xia
Haifeng Xia

approved: Zhengming Ding
Zhengming Ding
Chair

Jihun Hamm

Zizhan Zheng
Zizhan Zheng

Chen Chen
Chen Chen

Abstract

Deep neural networks have achieved promising performance on solving multiple computer vision tasks when trained on large-scale in-domain datasets. For some specific learning tasks, the model learned from the off-the-shelf well-labeled training set suffers from performance degradation when evaluated on the novel test set due to their distribution divergence. The practical dilemma motivates the emerging research topic named as "Unsupervised Domain Adaptation". This dissertation is centered with a novel perspective of cross-domain data access privileges to discuss various domain adaptation scenarios including unlimited cross-domain data access, source-data absent scenario and target-data missing scenario.

The first scenario is unsupervised domain adaptation where training model utilizes well-labeled source domain and unlabeled target samples. This condition is beneficial to learn domain-invariant representation with all available data. To mitigate domain shift, we propose a structural preserving generative method to perform graph alignment. Moreover, this thesis also considers a novel metric with cross-domain graph information to implement category-wise alignment.

The second scenario is source-free domain adaptation, where only a well-trained source model and target data are available for knowledge transfer and adaptation. To address this, we propose an adaptive adversarial network to improve classification ability and transfer source knowledge by developing a flexible target-specific classifier. To promote the model's robustness, the approach utilizes category-wise matching and self-supervised learning.

Finally, the target-data missing scenario comprises two sub-problems. The first sub-problem is when there are no target instances for learning the model. In this case, domain generalization (DG) can be used, which introduces multiple source domains to learn domain-invariant features. Furthermore, we extend DG by exploring the change of model generalization ability with imbalanced data distribution and use data augmentation to overcome it. The other strategy is zero-shot domain adaptation, which utilizes an additional task-irrelevant dataset to learn cross-domain contents and improve model generalization ability. Additionally, we formulate the second sub-problem as incomplete multi-view domain adaptation, where the multi-view source data and single-view target instances are available for model training. We adopt channel-wise change and enhancement to recover missing information and align various distributions.

Keywords: Domain Adaptation, Data Access Privileges, Knowledge Transfer

Visual Cross-Domain Adaptation Under Various Data Access Privileges

a dissertation
submitted on the twenty-sixth day of April, 2023
to the Department of Computer Science
of the School of Science and Engineering of
Tulane University
in partial fulfillment of the requirements
for the degree of
Doctor of Philosophy
by

Haifeng Xia

Haifeng Xia

approved: Zhengming Ding

Zhengming Ding
Chair

Jihun Hamm

Zizhan Zheng

Zizhan Zheng

Chen Chen

Chen Chen

Acknowledgment

First, I would like to express my sincere thanks to my academic advisor Prof. Zhengming Ding for his continuous guidance and unreserved support during the past four years. When I initially embarked on this challenging journey, his tremendous encouragement firstly helped me overcome my fear of language learning. After entering his laboratory, it was his persuasive suggestions and infinite enthusiasm that motive me to gradually understand how to do high-quality research work. While I expected to apply what I have learned to actual industrial product development, it was his generosity and openness that make my internships at MERL, NEC Lab and Google possible and successful. All in all, without Prof. Ding's help, this dissertation and my current achievements could not be possible. It is my honor and pleasure to work with Prof. Ding.

I would also like to thank my committee members, Prof. Jihun Hamm, Prof. Zizhan Zheng and Prof. Chen Chen for their valuable time, constructive comments and suggestions ever since my research prospectus. Their support and feedback provided a clearer research direction for me and better shaped my proposed research methods. Moreover, when I was in need on my faculty job searching, they are always willing to provide support. I am truly fortunate to have them as my committee members.

Finally, I would like to express my gratitude to my parents for their unconditional love and encouragement throughout the process of my study and many thanks to Muxi for her unfailing company and support during this particular journey. Besides, it is my pleasure to work and discuss with Taotao Jing on several research works.

Contents

1 Introduction 1

1.1 Background . . . 1

1.2 Related Works . . . 6

1.2.1 Unsupervised Domain Adaptation . . . 6

1.2.2 Source-Free Domain Adaptation . . . 7

1.2.3 Multi-Domain Learning . . . 7

1.3 Dissertation Organization . . . 8

2 Structural Generation for Unsupervised Domain Adaptation 10

2.1 Background . . . 10

2.2 The Proposed Algorithm . . . 12

2.2.1 Preliminaries and Motivation . . . 12

2.2.2 Cross-Domain Graph Alignment . . . 13

2.2.3 Source-Supervised Target Classifier . . . 15

2.2.4 Entropy Minimization . . . 17

2.2.5 Optimization . . . 18

2.3 Experiment . . . 20

2.3.1 Experimental Setting . . . 20

2.3.2 Comparison Results . . . 21

2.3.3 Ablation Study . . . 23

2.4 Conclusion . . . 25

3 Generative Metric Learning for Unsupervised Domain Adaptation 26

3.1 Background 26

3.2 The Proposed Method 29

3.2.1 Preliminaries and Motivation 29

3.2.2 Maximum Structural Generation Divergence 31

3.2.3 Overall Objective 36

3.3 Experiment 37

3.3.1 Benchmark Datasets 37

3.3.2 Experimental Setup 39

3.3.3 Comparison Results 40

3.3.4 Empirical Analysis 44

3.4 Conclusion 46

4 Dual-Classifier Adversarial Learning for Source-Free Domain Adaptation 47

4.1 Background 47

4.2 The Proposed Method 50

4.2.1 Preliminaries 50

4.2.2 Adaptive Adversarial Network 51

4.2.3 Overall Objective and Optimization 55

4.3 Experiments 56

4.3.1 Experimental Details 56

4.3.2 Comparison Results 57

4.3.3 Empirical Analysis 59

4.4 Conclusions 62

5 Representation Generation for Imbalanced Domain Generalization 63

5.1 Background 63

5.2 The Proposed Method 66

5.2.1 Motivation and Preliminaries 66

5.2.2 Balanced Generative Paradigm 68

5.2.3 Generative Inference Network (GINet) 69

5.2.4 Theoretical Analysis 73

5.3 Experiments 74

5.3.1 Experimental Setup 74

5.3.2 Comparison Results 77

5.3.3 Empirical Analysis 80

5.4 Conclusions 83

6 Representation Decomposition for Zero-Shot Domain Adaptation 84

6.1 Background 84

6.2 The Proposed Method 87

6.2.1 Preliminaries and Motivation 87

6.2.2 Adaptive Feature Separation 88

6.2.3 Hybrid Generation 91

6.2.4 Training and Inference 92

6.3 Experiments 93

6.3.1 Datasets and Comparisons 93

6.3.2 Implementation Details 94

6.3.3 Experimental Results 95

6.3.4 Ablation Study 97

6.4 Conclusion 100

7 Representation Fusion for Incomplete Multi-View Domain Adaptation 101

7.1 Background 101

7.2 Proposed Method 104

7.2.1 Preliminary & Motivation 104

7.2.2 Source Channel Enhanced Network 105

7.2.3 Adaptive Knowledge Transfer Network 107

7.2.4 Overall Objective 109

7.2.5 Theoretical Analysis 110

7.3 Experiments 111

7.3.1 Experimental Details 111

7.3.2 Comparison of Results 113

7.3.3 Empirical Analysis 114

7.4 Conclusion 116

8 Conclusion & Future Works **118**

References **120**

Chapter 1

Introduction

1.1 Background

The past decade witnesses the thriving development of the high-tech industry, especially the Internet which allows customers to upload and share their daily images with others through different media [1, 2]. Meanwhile, technical advance on visual or signal sensors makes it possible and affordable for us to collect considerable data [3, 4]. This situation naturally stimulates demands to understand contents from the corresponding data modality and further utilize the learned knowledge to provide clients with superior service. Fortunately, the continuous exploration of deep learning technique is gradually disintegrating this practical difficulty and has obtained the appealing progress on overcoming several computer vision tasks such as image classification [5, 6], object detection [7, 8] and semantic segmentation [9, 10]. These successful applications are mainly attributed to its powerful fitting ability by extracting the hierarchical semantics from the original inputs and using them to deduce the decision.

To obtain such an ability, the training of deep neural network (DNN) generally requires accessing large-scale label-sufficient datasets such as ImageNet [11] and PASCAL VOC 2008 [12]. However, the manual annotation on each instance becomes time-consuming and expensive in many real-world application scenarios [13]. For example, in the autonomous driving field, machines need to perceive environmental change and accurately distinguish between pedestrians, vehicles and traffic lights. It also indicates that each object of image or video is supposed to be annotated, which is a heavy burden for product developers. This drawback obviously obstructs the popularization of the deep learning paradigm in industrial manufacturing. On the other hand, the effective deployment of the model learned with supervised learning is built on the strict assumption that training and test samples belong to the identical distribution [14, 15]. However, realistic working situations hardly satisfy this point due to the varying captured environments or the difference of sensors. For instance, during daytime and nighttime, a camera in a vehicle captures two entirely different images at the same location. Under this condition, training and test images are collected from two distributions or domains. The former is typically regarded as source domain, while the latter is named as the target domain. The distribution difference of their input data is also formally defined as domain shift easily triggering the significant performance degradation when evaluating the well-learned source model on target domain [16,17]. Hence, surmounting the mentioned application bottlenecks attracts more attentions in machine learning and computer vision communities in recent years.

On the road of exploration, transfer learning casts the light in the darkness and becomes a proven strategy. In fact, this learning paradigm is imitating human cognitive procedure where ones can learn knowledge by observing few demonstrations and apply the off-the-shelf information to solve new situations. Similarly, pre-training the deep model over the well-labeled source domain can be analogized as the knowledge accumulation stage and then further adapting the well-learned source model into the related yet different target distribution is conducting the second evolutionary process. In this case, the adapted DNN is likely to be high-generalization on identifying test samples collected from the target distribution. In a nutshell, transfer learning is a procedure of transferring and adapting source knowledge to solve the similar task in target domain.

Along with this direction, [18] theoretically analyzes the upper bound of target error and points out that the elimination of cross-domain feature distribution difference assists model in achieving the optimal knowledge transfer. With the guidance of theorem, the mainstream solutions [19, 20] present unsupervised domain adaptation (UDA) setting. Specifically, for image classification task, source and target images are sampled from different distributions but share the identical categorical space, and cross-domain samples are both available for model training in UDA. The loose cross-domain data accessibility allows these domain adaptation algorithms to explicitly measure the discrepancy of their feature representations and progressively mitigate it to align various distributions via metric learning [21] or generative adversarial manner [22]. Their empirical studies also fully suggest that the knowledge transfer mechanism can evidently promotes model performance in target domain by utilizing both source data with supervision and unlabeled target instances.

Although these UDA methods manifest an appealing adaption ability, deploying them into realistic industrial applications still encounters large obstructions, especially for dealing with sensitive data or finite-storage devices. For example, to improve disease diagnosis techniques, different hospitals normally build collaboration to train a precise diagnosis system. Concretely, without sufficient resources, the newly-built medical institutions difficultly conduct the expensive manual annotation on the collected samples, while the other prestigious hospitals have already accumulated abundant well-labeled diagnosis records. Considering the difference in their medical devices, transferring and adapting knowledge from the prestigious hospitals to the newly-built ones benefits and accelerates the system construction of the latter one. However, medical diagnosis data involves the sensitive personal information of patients and is always protected at the local hospital to avoid privacy leakage. This also means that these prestigious hospitals fail to follow the conventional UDA policy to share their patients' data for domain adaptation. Besides, in terms of autopilot, the vehicle is always equipped with lightweight (limited storage and computing space) perception and recognition modules to reduce load and response time. Thus, it is unrealistic to reserve lots of source images in these components for model adaptation. In summary, these two practical cases are both indicating that source data is prohibited from appearing in the target model learning for many real-world applications. This observation triggers one most straightforward question: *How to transfer and adapt source knowledge to the target domain without accessing source data?*

Apart from source-data absent situation, there still exist several real-world scenarios which it is challenging for UDA methods to overcome. The main cause of difficulty stems from the strict basic assumption that learning domain-invariant representation needs to access both source and target images. But, it is not uncommon to discover that the collection of samples from the unseen target distribution tends to be tough and impossible. For instance, in the process of solving criminal cases, the police can usually draw the sketch image of suspects based on the testimony of eyewitnesses and match them with other RGB images captured by public surveillance cameras. Due to the distinction of image style, the simple and direct matching likely reduces the accuracy of detection. Definitely, UDA based solutions easily mitigate the negative influence of domain shift by regarding sketch and colorful images as source and target samples respectively. Importantly, the colorful images of suspects are generally unavailable for training the recognition system. The conflict naturally triggers the other research question: *How to learn a high-generalization model for the unseen target distribution by only training it with the available source data points?* For this problem, one recent exploration is domain generalization [23, 24] utilizing multiple source domains to learn domain-invariant features and directly generalizing the model into the unseen target domain. This learning strategy heavily depends on multi-source data quality. Specifically, the imbalance of category distribution across multiple source domains negatively affects the model generalization. Hence, an urgent sub-problem to be solved is that *can we effectively generalize well-learned knowledge into the unseen target domain with imbalanced source data distribution*? Back to the target-data absent scenario, when there is prior knowledge about the unseen target image modality, *can we introduce another additional multi-modality dataset to assist knowledge generalization?* Concretely, for the mentioned recognition system, we can select one existing dataset with the paired sketch and RGB images and utilize them to assist model removing modality-relevant information. With the auxiliary feature learning, DNN focuses on capturing the important semantics of objects and generalizing these contents to identify the unseen target images.

Additionally, the standard UDA setting supposes that source or target image is only captured by a single sensor such as normal camera or depth one. But many industrialized products like autonomous vehicles install multiple sensors in the new version to boost model performance with multi-view contents. Thus, a few pioneers [25, 26] have explored multi-view domain adaptation (MVDA), where source and target data are both collected from multiple sensors. The intuitive idea is to convert MVDA into a UDA problem by independently aligning source and target instances within each view and fusing multi-view semantic information within individual domain. They have achieved promising performance on solving MVDA and abundant empirical studies illustrate that the simple alignment-and-fusion promotes model performance on identifying target samples with more enriched data collected by multiple sensors. However, equipment rehabilitation to upgrade previous single-sensor devices with multiple sensors causes additional cost overhead, which makes MVDA to be invalid for several practical application scenarios. Instead, we post a question that *"Can we develop more effective domain adaptation algorithms to benefit single-sensor target data from enriched source data with multiple sensors?"*. In other words, this problem suggests that there are multi-view source samples and single-view instances in the target domain. The exploration of this problem is highly demanded since it can efficiently solve the compatibility of the system after a product upgrade.

Highlight of research topics

In this dissertation, our main expectation is to make domain adaptation methods applicable to these real-world industrial scenarios with different data access privileges. Our research starts from the conventional unsupervised domain adaptation without limitations on cross-domain data accessibility. And then we explore source-data absent and target-data absent knowledge transfer scenarios. Moreover, the demand of the autonomous driving field inspires us to explore incomplete multi-view domain adaptation. Hence, this thesis mainly tackled the following research questions:

1. How can we conduct more effective knowledge transfer to solve the UDA problem?
2. Can we transfer and adapt source knowledge to the target domain without accessing source data during the process of model adaptation?
3. How to generalize well-learned knowledge into the unseen target domain under imbalanced source data distributions or with the auxiliary multi-modality dataset?
4. Can we develop more effective domain adaptation algorithms to benefit single-sensor target data from enriched source data with multiple sensors?

1.2 Related Works

1.2.1 Unsupervised Domain Adaptation

The current unsupervised domain adaptation methods mainly includes two categories: metric-based algorithms and generative adversarial feature learning. Existing metric-based methods evaluate distribution difference from three perspectives: domain-level, class-level and sample-level. Considering the inaccessibility of target labels, domain-level alignment generally explores statistics of distribution to measure cross-domain discrepancy. Specifically, MMD [27] as the most popular and effective tool attempts to measure and narrow the distance of their first-order moment (mean value). Deep Adaption Network (DAN) [21] extends MMD into multi-kernel formulation on top layers of neural network. To capture more properties of distribution, deep coral [28] focuses on the second-order statistic (co-variance) to compare cross-domain difference. Along this side, [29] exploits the consistence of high order central moments to align various distributions. Benefiting from the application of pseudo label, class-level alignment attracts more attentions. The pseudo annotation is generally from the structural prediction or the output of network. Concretely, CAN [30] firstly assigns pseudo labels to target samples and then reformulates MMD by introducing cross-domain intra-class and inter-class concepts. To improve the quality of pseudo labels, [31] explores the label propagation to refine the label generation from feature level. In addition, [32] expects to learn similar representation for source and target samples from the same category with the incorporation of class prior probability and MMD metric. Similarly, [33] adopts clustering fashion to align source and target class centers. Besides, sample-level metric is developed to explore domain shift. For instance, ETD [34] considers cross-domain sample correlation as weight to adjust optimal transport constraint.

Different from metric-method to explicitly calculate distribution discrepancy, generative domain adaption borrows the spirit of generative adversarial network (GAN) [35] to learn domain-invariant feature representation. Generally, such a network architecture includes two primary components: feature generator and domain discriminator. The feature generator attempts to fool the domain discriminator with cross-domain features until it fails to identify which domain they come from. Along this line, DANN [36] develops one domain classifier to identify various distributions and explores gradient reversal layer to adjust cross-domain feature learning. Inspired by the merits of conditional GAN [37], CDAN [38] exploits linear combination manner to associate feature with the predicted class distribution and regards the connection as input for discriminator. Moreover, [22] utilizes the inconsistent prediction of two classifiers to distinguish source features from target ones and considers them as one discriminator to obtain domain-invariant representations in adversarial manner. To balance transferability and discriminability, DADA [39] incorporates classifier and discriminator into a unified framework. In addition, [40] adopts GAN to stylize source images with target semantic and take them as intermediate domain to connect source and target domains. Similarly, DM-ADA [41] produces the additional images in the intermediate domain via the random combination of cross-domain images and progressively achieve distribution match.

1.2.2 Source-Free Domain Adaptation

Source-free domain adaptation provides the well-trained source model and target data without any access to source data during the training stage. Under this condition, the conventional UDA methods become invalid since they fail to match source feature with target ones. Most recently, [42–44] discover that the original trained model conceals lots of knowledge of source feature distribution. Thus, with the supervision of source classifier, [42, 43] attempt to produce novel target samples closer to source domain and then align the novel and original target instances in high-level features via adversarial manner. Similarly, [44] freezes the source classifier and applies pseudo-label to optimize feature generator, which aims to move target features into the unseen source feature domain. However, the significant domain discrepancy or imbalanced distribution of source domain has a negative influence on the generalization of source model, which increases the difficulty of adapting target feature to source classifier.

1.2.3 Multi-Domain Learning

Domain Generalization (DG) as a more challenging task only accesses to multi-source datasets without any prior knowledge from target domain [45, 46]. The intuitive attempts

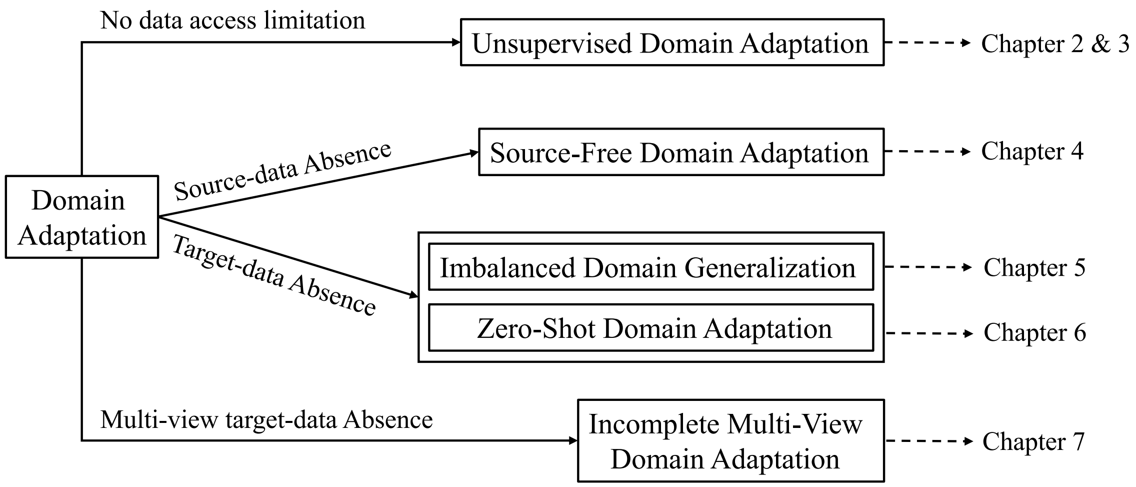

No data access limitation
Unsupervised Domain Adaptation
Chapter 2 & 3
Source-data Absence
Source-Free Domain Adaptation
Chapter 4
Domain Adaptation
Target-data Absence
Imbalanced Domain Generalization
Chapter 5
Zero-Shot Domain Adaptation
Chapter 6
Multi-view target-data Absence
Incomplete Multi-View Domain Adaptation
Chapter 7

Figure 1.1: Main research questions explored in this dissertation.

consider simulating the target images via the combination of multi-source instances to improve the diversity of input data [47–49]. Another effective direction aims to learn domain-invariant features from multi-source domains [50]. Concretely, the low-rank parameterized CNN structure [51] expects to capture intrinsic attribution from various and complicated visual signals. Due to the success of Gradient Reversal Layer (GRL) component, [50] adopts adversarial learning manner with the guidance of one discriminator to learn the generalized representations. Motivated by episodic training, [52] seeks for improving the robustness of model through the gradual exposure between network and new domain. Recent works [53, 54] inspired by jigsaw puzzle excavate the inherent information of object via the recognition of relationship among image patches. Unlike them, this paper considers the negative influence of imbalanced data scale across source domains and categories on learning better-generalized model for IDG scenario.

1.3 Dissertation Organization

As Figure 1.1 shows, this thesis mainly solves domain adaption with four different constraints on data accessibility, i.e., free cross-domain data access, source-data or target-data absent scenarios and incomplete multi-view domain adaptation. Hence, the rest of this dissertation is organized as follows.

In Chapters 2 & 3, to surmount the limitation of existing UDA methods on achieving distribution alignment, we attempt to utilize intrinsic cross-domain structural information to instruct the domain-invariant feature learning. Specifically, the structural knowledge is used to construct an intermediate domain associating source and target domains to gradually eliminate their domain shift via metric learning and adversarial mechanism.

In Chapter 4, we mainly explore the source-free domain adaptation problem where the knowledge adaptation on the target domain only accesses the well-learned source model and unlabeled target instances. Without the assistance of source data, we adopt a dual-classifier to distinguish source-similar samples from source-dissimilar ones and gradually align them via adversarial and contrastive constraints.

In Chapters 5 & 6, we consider the target-data absent scenario where samples from target distribution are invisible for the entire model training and adopt two learning strategies to address it. One is generating domain-invariant representation over multiple source domains to promote the generalization of the model. The other one instructs the model to decompose features into task-relevant semantics and task-irrelevant ones by training it on an additional multi-modality dataset. In this case, the model can gradually extract and generalize intrinsic source knowledge.

In Chapter 7, we discuss incomplete multi-view domain adaption where source samples are captured by multiple views (sensors) while target instances are described by one single view. The main challenge is how to effectively transfer sufficient multi-view semantic information to benefit the task on target domain. For this, we adopt channel-wise exchange mechanism and optimal transport to fulfill representation fusion and knowledge transfer.

In Chapter 8, we conclude the main research questions explored in this dissertation and our proposed methodologies for domain adaptation with different data access privileges. Moreover, we briefly illustrate our future potential research topics along with the footprint of current exploration.

Chapter 2

Structural Generation for Unsupervised Domain Adaptation

2.1 Background

Deep neural networks have achieved an increasing number of successes in computer vision community with a great deal of well-labeled data, which allows deep learning models to easily capture abstract and complex relationship between feature and category [22]. In reality, however, collecting abundant data with annotation becomes too difficult and expensive in many learning tasks. The intuitive motivation to address the realistic issue is to apply knowledge extracted from model trained with available annotated samples into target tasks. Such a strategy frequently tends to be vulnerable for the problem of domain shift [55] as the trained model is more likely to be invalid when assessed on unlabeled target domain having various distribution with training source. Specifically, for visual data, domain shift results from distinctions of light condition occlusions and background [56].

Unsupervised Domain Adaptation (UDA) is a promising technique to train a model obtaining lower risk when evaluated on target domain [57–61]. Existing UDA methods [38, 62, 63] generally minimize the risk on source data firstly and then employ appropriate statistical property to eliminate cross-domain discrepancy. There are two common manners to measure discrepancy between distributions of two domains, i.e., discrepancy measurement

[21, 27] and domain adversarial confusion [22, 64]. Specifically, discrepancy measurement like maximum mean discrepancy employs statistical indication (mean of distribution) to measure cross-domain difference and aligns the distribution of two domains by constraining this indication. While domain adversarial confusion aims to seek a domain invariant feature generator for both domains with a domain confusion discriminator in an adversarial training manner. However, these methods still remain restrictive in the alignment between feature and category due to the neglect of class-level information [65]. They generally suffer from two challenging issues: 1) mis-alignment of cross-domain samples from various classes and 2) the learned classifier would lack of generalization on target domain [30].

To alleviate these disadvantages, target pseudo labels are introduced to effectively enhance class-level alignment during the training process [66,67]. Moreover, [32] considers the class prior probability defined on two domains as class-specific weight and modifies original MMD with auxiliary weights to promote discriminative ability of classifier for target domain. Similarly, a novel metric measure formulated in [30] includes intra-class domain discrepancy and inter-class domain discrepancy. On the other hand, recent studies [68, 69] pay more attention to the second issue, which attempts to make the learned decision boundary robust for target domain. The common strategy to address this challenge designs two domain-specific classifiers. Subsequently, [70] regards two classifiers as various views for the same samples of source domain and maximum their distinction to learn a robust classifier for samples from target domain. In addition, [71] develops sliced wasserstein discrepancy (SWD) connecting feature distribution alignment and wasserstein metric to promote the discrimination of target classifier. However, training a target-specific classifier with samples from corresponding domain is inaccessible, which certainly obstructs classification accuracy. This issue stems from the inaccessibility of target label.

In this paper, we propose a Generative cross-domain learning via Structure-Preserving (GSP) model to incorporate samples of target domain into training phase with source supervision (Fig. 2.1). Specifically, a novel metric discrepancy is defined to measure cross-domain distinction in terms of the topological structure including information of node and edge. In order to minimize cross-domain discrepancy, two-level alignments (i.e., edge-level and node-level) are designed to enhance the mitigation of domain mismatch. The edge-level alignment aims to discover matching relationship between two domains according to node and degree, while the node-level alignment exploits learned matching relationship to restrict feature representation across two domains. Moreover, we develop a source-supervised target classifier which supervises feature learning of target domain with source label. Furthermore, we adopt a symmetrical and adversarial manner to train two domain-specific classifiers, which not only maximize the difference between two classifiers but also extract effective domain invariant features. To this end, our contributions are summarized as following:

- We introduce a novel metric measure in terms of graph distribution and formulate alignments of node-level and edge-level. The edge-level alignment is employed to extract cross-domain matching relation, while node-level operation aims to align feature representation.
- To promote the discriminative ability of classifier, we develop source-supervised target classifier fed with the combination of matching relation and features from target domain. Moreover, we apply symmetric adversarial manner to train two domain-specific classifiers.
- We evaluate our proposed model (GSP) on several visual cross-domain benchmarks. GSP approach outperforms competitive methods in most domain adaptation tasks, demonstrating the effectiveness of solving UDA problem. Extensive analysis illustrates the function of each component in GSP method.

2.2 The Proposed Algorithm

2.2.1 Preliminaries and Motivation

For UDA, we are generally given source dataset $\mathcal{D}_s = \{(x_i^s, y_i^s)\}_{i=1}^{n_s}$ and target dataset $\mathcal{D}_t = \{x_i^t\}_{i=1}^{n_t}$ where $\mathcal{D}_s$ includes n_s data samples $\{x_i^s\}_{i=1}^{n_s}$ with its corresponding label set $\{y_i\}_{i=1}^{n_s}$, and $\mathcal{D}_t$ consists of n_t data instances $\{x_i^t\}_{i=1}^{n_t}$ and the label information for target domain is unknown. Although it is obvious that the same label space is shared by these two domains, the distributions of their data sample sets are different, which limits the performance of the trained model from source to target domain. Minimizing the source risk and bounding the discrepancy between two various distributions effectively improve the performance of model, which has been verified by abundant theoretical analyses.

In this work, we rethink UDA problem from perspective of graph distribution and propose a novel generative model with structure preserving. Concretely, samples within each domain constitute graph structure with information of node, edge and degree. Although there is distribution discrepancy across two domains, topological structures of them are more likely to be similar. Thus, the proposed method matches topological information across two domains through Gromov-Wasserstein (GW) discrepancy [72] defined over graph and leverages the learned relationship to eliminate discrepancy between $\mathcal{D}_s$ and $\mathcal{D}_t$ with cross-domain graph alignment. In addition, we develop a novel source-supervised target classifier jointly with cross-domain alignment to make the trained classifier robust to unlabeled target learning.

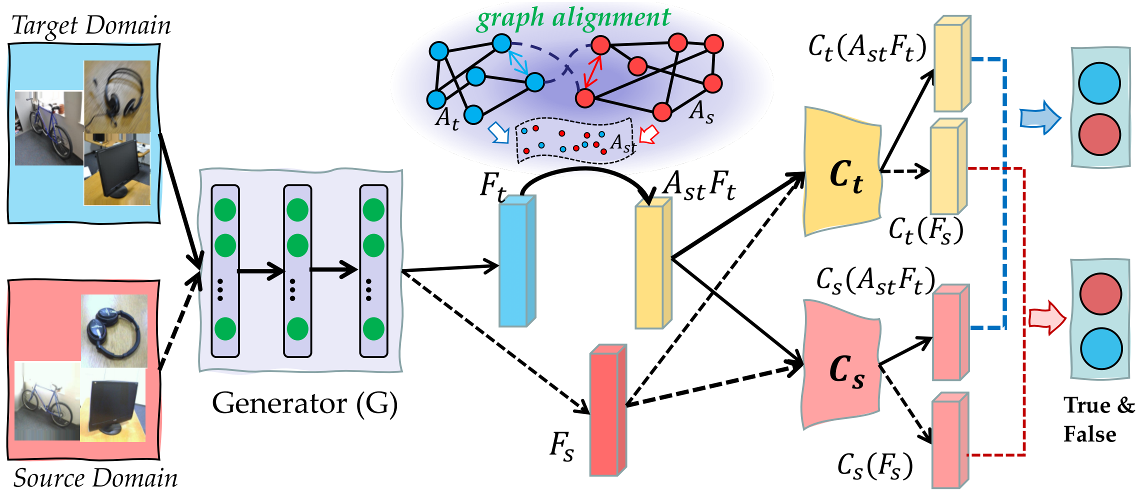


Figure 2.1: Overview of the proposed architecture, where features F_s and F_t are extracted from raw data through generator (VGG or ResNet), and then we capture matching relationship (blue dotted line) of two domains according to graph distribution. Moreover, two classifiers are built and fed with same input. We adopt domain adversarial training manner to maximum the difference between them.

2.2.2 Cross-Domain Graph Alignment

Existing approaches [73, 74] achieve promising performance by benefiting from deep neural networks, e.g., VGG [75] and ResNet [76]. Those algorithms explore existing deep neural networks as backbone to extract general feature representation and stack cross-domain alignment at the top. Suppose $F_s = \{f_i^s\}_{i=1}^{n_s}$ and $F_t = \{f_j^t\}_{j=1}^{n_t}$ are feature representations from two domains $\mathcal{D}_s$ and $\mathcal{D}_t$, respectively. With extracted features, we define the measurable graphs of source domain and target domain as $G_s(\mathcal{V}_s, A_s, p_s)$ and $G_t(\mathcal{V}_t, A_t, p_t)$, where $\mathcal{V}_s = \{v_i\}_{i=1}^{n_s}$ $(\mathcal{V}_t)$ is the set of nodes in the corresponding domain, the similarity or distance between elements in source domain (target domain) is denoted as $A_s = [a_{ij}^s] \in \mathbb{R}^{n_s \times n_s}$ (A_t), and $p_s(p_t)$ represents Borel probability measurement defined on $\mathcal{V}_s(\mathcal{V}_t)$. In practice, p_s (p_t) represents empirical distribution of nodes and it is estimated by normalized node degree.

To effectively match two different domains, we propose two-level cross-domain alignments, i.e., node-level and edge-level. First of all, we explore GW distance to measure the edge similarity across two domains [77]. Metric measures of source domain and target domain are defined as d_s, d_t, respectively. In term of these definitions, we extend GW method to measure the discrepancy of cross-domain topology structure and have the following formulation of edge-level alignment $\mathcal{L}_e$:

$$\begin{aligned}\mathcal{L}_e &= \Big(\sum_{i,j\in\mathcal{V}_s}\sum_{i^{'},j^{'}\in\mathcal{V}_t}|A_{ij}^s - A_{i^{'}j^{'}}^t|A_{i,i^{'}}^{st}A_{j,j^{'}}^{st}\Big)^{\frac{1}{p}} \\ &= \langle L(A_s, A_t, A_{st}), A_{st}\rangle,\end{aligned} \tag{2.1}$$

where $A_{st} = \{A_{st} \in \mathbb{R}_+^{n_s \times n_t} | A_{st}\mathbb{1}_{n_t} = p_s, A_{st}^{\mathrm{T}}\mathbb{1}_{n_s} = p_t\}$ is the joint distribution of node degree, i.e., $A_{st} \in \Pi(p_s, p_t)$, $L(A_s, A_t, A_{st}) = A_s p_s \mathbb{1}_{n_t}^{\mathrm{T}} + \mathbb{1}_{n_s} p_t^{\mathrm{T}} A_t^{\mathrm{T}} - 2A_s A_{st} A_t^{\mathrm{T}}$ is derived from [78], and $\langle A, B\rangle$ is the inner product of matrices A and B.

To further mitigate the domain mismatch, we bridge the node-level domain gap. In practice, v_i^s (v_j^t) can be represented by the feature f_i^s (f_j^t). Targeting at coupling the relationship between features from various domains, we further exploit the learned structured information to constrain feature representation and reduce discrepancy of two domains. In addition, A_{ij}^{st} also indicates the probability that v_i^s and v_j^t belong to the same category. Thus, we define the node-level alignment as $\mathcal{L}_n$:

$$\mathcal{L}_n = \|F_s - A_{st}F_t\|_{\mathrm{F}}^2, \tag{2.2}$$

where $\|\cdot\|_{\mathrm{F}}$ is the Frobenius norm.

To sum up, our two-level cross-domain graph alignment module is defined by incorporating Eq. (2.1) and (2.2) together as follows:

$$\mathcal{L}_g = \mathcal{L}_e + \mathcal{L}_n. \tag{2.3}$$

Remark: Edge-level alignment in Eq.(2.1) integrates the distinction between arbitrary edges from various domains and graphs' degree information into a single system. The distance of cross-domain edge reflects domain discrepancy embedded into A_{st}. Optimal A_{st} explores a probabilistic assignment to match the source nodes to the target ones. Compared to edge-level alignment, node-level alignment directly focuses on feature representation. A_{ij}^{st} indicates the probability that the source feature f_i^s and target feature f_j^t belong to the same category. According to Eq. (2.2), cross-domain samples with the same label tend to be clustered in the shared space with similar feature representation.

2.2.3 Source-Supervised Target Classifier

Due to the lack of label information in target domain, existing methods to solve UDA problem only employ samples from source domain to train a domain-invariant classifier shared by target domain. Other works [70, 79] alternatively design two classifiers corresponding to two domains and maximize distinction of them. To enhance the generalization ability of the classifiers to target samples, existing works normally explore pseudo labels by involving the target supervision iteratively [66, 67]. However, the fundamental challenge (e.g., to learn a robust classifier for target domain) is still unsolved as ground-truth target label is not accessible. In order to address this issue, we develop a novel source-supervised target classifier $C_t(\cdot)$ with structure preserving, as well as a traditional source-supervised classifier $C_s(\cdot)$ under a symmetric adversarial training manner.

We firstly introduce how to feed unlabeled target samples into the source-supervised target classifier and then present the whole symmetric adversarial architecture. As discussed in section 2.2.2, features F_s extracted from $\mathcal{D}_s$ can be represented by features of target domain F_t under node-level alignment, i.e., $\|F_s - A_{st}F_t\|_{\mathrm{F}}^2$. Without loss of generality, arbitrary f_i^s has the formulation $f_i^s \approx \sum_{j=1}^{n_t} a_{ij}^{st} f_j^t$. The larger a_{ij}^{st} not only demonstrates v_i^s has similar topological structure with v_j^s but also indicates f_i^s and f_j^t come from the same class. This strategy is also considered as a tool extracting samples with larger a_{ij}^{st} from target domain and ignoring influence of other samples to code f_i^s. Most likely, the selected samples share the same label with f_i^s, and are input to train the classifier, which dramatically promote the discriminative ability of classifier for samples in target domain.

Thus, C_s and C_t are developed by taking $\{F_s, Y_s\}$ and $\{A_{st}F_t, Y_s\}$ as input, respectively. Noted that $A_{st}F_t$ shares the same label information with F_s. C_t also learns to identify the interface among various classes in source domain. Interestingly, $C_t(\cdot)$ trained on $A_{st}F_t$ should also be valid to recognize F_t, since $A_{st}F_t$ and F_t share the same feature space. In this sense, we obtain the target classifier with ground-truth source supervision by transforming the target features into source ones. Note that $A_{st}F_t$ can be treated as a bridge to gap the source and target domains.

However, considering that the task of C_t is to trigger more accurate predictions on target domain, the probabilities generated from $C_t(A_{st}F_t)$ and $C_t(F_s)$ should become different. Inspired by [22], symmetric adversarial architecture is exploited to achieve this goal. From Fig. 2.1, there are two parallel classifiers C_s and C_t sharing the same input F_s and $A_{st}F_t$. And C_s and C_t are built in the same architecture including Fully-Connected (FC) layers and one Softmax layer. For an arbitrary feature input such as f_i^s, the output of C_s and C_t are denoted as $q_s(f_i^s) \in \mathbb{R}^C (q_s \mathbb{1}_C = 1)$ and $q_t(f_i^s) \in \mathbb{R}^C (q_t \mathbb{1}_C = 1)$, where C is the number of classes.

Given features F_s and $A_{st}F_t$, two classifiers generate four types of probabilities: $q_s(F_s)$, $q_s(A_{st}F_t)$, $q_t(F_s)$ and $q_t(A_{st}F_t)$. We train C_s and C_t to make prediction for any input by minimizing the following cross-entropy loss:

$$\begin{aligned}\mathcal{L}_s = -\frac{1}{n_s}\Big(&\sum\nolimits_{i=1}^{n_s} y_i^s \log\big(q_s(f_i^s)\big)\\ &+ \sum\nolimits_{i=1}^{n_s} y_i^s \log\big(q_s(\sum\nolimits_{j=1}^{n_t} a_{ij}^{st} f_j^t)\big)\Big),\\ \mathcal{L}_t = -\frac{1}{n_s}\Big(&\sum\nolimits_{i=1}^{n_s} y_i^s \log\big(q_t(f_i^s)\big)\\ &+ \sum\nolimits_{i=1}^{n_s} y_i^s \log\big(q_t(\sum\nolimits_{j=1}^{n_t} a_{ij}^{st} f_j^t)\big)\Big).\end{aligned} \tag{2.4}$$

Although C_s and C_t leverage same features as input, they should have various identifying functions. The primary purpose of C_s is to improve prediction accuracy of feature F_s while C_t pays more attention to the prediction of $A_{st}F_t$. To achieve this goal, we extract feature $H_s(H_t)$ from classifier $C_s(C_t)$ before the Softmax layer and then concatenate features into $H_{st}^s = [H_s(F_s), H_t(F_s)]$ and $H_{st}^t = [H_s(A_{st}F_t), H_t(A_{st}F_t)]$. Subsequently, softmax operation is applied to obtain probability distribution $[q_s^*(F_s), q_t^*(F_s)]$ and $[q_s^*(A_{st}F_t), q_t^*(A_{st}F_t)]$. Alternatively, $q_s^*(F_s)$ should be larger than $q_t^*(F_s)$ but $q_s^*(A_{st}F_t)$ is supposed to have smaller value than $q_t^*(A_{st}F_t)$. We adopt the domain adversarial training manner in [22] by minimizing the following additional cross-entropy losses:

$$\begin{aligned}\mathcal{L}_{s_a} &= -\frac{1}{n_s}\sum\nolimits_{i=1}^{n_s}\log(\sum\nolimits_{k=1}^{C} q_{s_k}^*(f_j^s)),\\ \mathcal{L}_{t_a} &= -\frac{1}{n_s}\sum\nolimits_{i=1}^{n_s}\log(\sum\nolimits_{k=1}^{C} q_{t_k}^*(\sum\nolimits_{j}^{n_t} a_{ij}^{st} f_j^t)).\end{aligned} \tag{2.5}$$

To this end, we can integrate Eq. (2.4) and Eq. (2.5) into the following Eq. (2.6) to train classifiers by minimizing:

$$\mathcal{L}_c = \mathcal{L}_s + \mathcal{L}_t + \mathcal{L}_{s_a} + \mathcal{L}_{t_a}, \tag{2.6}$$

Thus, this loss function involves classification task and domain adversarial task.

2.2.4 Entropy Minimization

Although source-supervised target classifier leverages collaboration of target samples to improve discrimination of classifier, there is no chance for target classifier to access features of target domain directly. To avoid this issue, we adopt Entropy minimization (EM) method widely used in [80] to promote the robustness of classifier. Entropy minimization function aims to simultaneously optimize two classifiers and has the following formulation:

$$\begin{aligned}\mathcal{L}_{em} = &-\frac{1}{n_t}\sum\nolimits_{i=1}^{n_t} q_s(f_i^t)\log(q_s(f_i^t))\\ &-\frac{1}{n_t}\sum\nolimits_{i=1}^{n_t} q_t(f_i^t)\log(q_t(f_i^t)),\end{aligned} \tag{2.7}$$

where $q_s(f_j^t)$ indicates the probability of target sample f_j^t and $q_t(f_j^t)$ means the output of target classifier for f_j^t. During the initial training phase, features of target domain lacking of discrimination are simply labeled with incorrect category and are difficult to be identified correctly in the later training phase. According to suggestion in [22], we only employ entropy minimization loss function to train generator instead of updating all parameters in our network.

2.2.5 Optimization

There are three components: generator, graph alignment and classifier in our proposed model to be optimized iteratively. We provide the following four steps to illustrate the optimization.

Step A: During the initial training phase, we use source instances with corresponding label to train C_s and C_t and update generator G. Although such a simple training manner is difficult to address domain shift problem, generator to some extent learns discriminative features for two domains. In terms of these extracted features, we can calculate cosine distance within each domain as A_s and A_t and then obtain the cross-domain similarity to initialize A_{st}.

Step B: The classifier C_t trained in the first phase produces pseudo label $\hat{Y_t}$ for target domain X_t. We then calculate a mask matrix $\mathcal{M} = Y_s\hat{Y_t}^{\mathrm{T}}$ to filter the irrelevant elements of A_{st} with the formulation as $\mathcal{M} \odot A_{st}$, where $\odot$ means element-wise product operation. Subsequently, we optimize A_{st} according to Eq. (2.3) and learn optimal cross-domain graph matching relation.

Step C: In this step, we train two classifiers C_s and C_t when fixing generator G. We take F_s and $A_{st}F_t$ as input both with source labels as supervised signal. In addition, classifier loss not only achieves classification task but also minimizes domain adversarial loss. Under this condition, classifiers are updated according to:

$$\min_{C_s,C_t} \mathcal{L}_s + \mathcal{L}_t + \mathcal{L}_{s_a} + \mathcal{L}_{t_a}. \tag{2.8}$$

Step D: Due to symmetric adversarial training, generator should confuse classifiers with $A_{st}F_t$ and F_s. Concretely, target classifier considers F_s as true while source classifier produces more value for input $A_{st}F_t$. Thus, we define a domain loss as $\mathcal{L}_d =$

Table 2.1: Top-1 Accuracy (%) on Office-31 dataset for UDA (ResNet-50) and the best result is in bold type.

Method	ResNet-50	DNN	DANN [36]	JAN [27]	SimNet [81]	SymNets [22]	TADA [82]	SAFN [83]	Ours
A→W	68.4	80.5	82.0	85.4	88.6	90.8	**94.3**	90.3	92.9
D→W	96.7	97.1	96.9	97.4	98.2	**98.8**	98.7	98.7	98.7
W→D	99.3	99.6	99.1	98.4	99.7	**100**	99.8	**100**	99.8
A→D	68.9	78.6	79.7	77.8	85.3	93.9	91.6	90.7	**94.5**
D→A	62.5	63.6	68.2	69.5	73.4	74.6	72.9	73.4	**75.9**
W→A	60.7	62.8	67.4	68.9	71.6	72.5	73.0	71.2	**74.9**
Avg	76.1	80.4	82.2	82.9	86.2	88.4	88.4	87.6	**89.5**

Table 2.2: Top-1 Accuracy (%) on Office-Home dataset for UDA (ResNet-50) and the best result is in bold type.

Method	ResNet-50	DANN [36]	JAN [27]	DSR [84]	SymNets [22]	TADA [82]	SAFN [83]	Ours
Ar→ Cl	34.9	45.6	45.9	53.4	47.8	53.1	52.0	**56.8**
Ar→ Pr	50.0	59.3	61.2	71.6	72.9	72.3	71.7	**75.5**
Ar→ Rw	58.0	70.1	68.9	77.4	78.5	77.2	76.3	**78.9**
Cl→ Ar	37.4	47.0	50.4	57.1	**64.2**	59.1	**64.2**	61.3
Cl→ Pr	41.9	58.5	59.7	66.8	71.3	71.2	**69.9**	69.4
Cl→ Rw	46.2	60.9	61.0	69.3	74.2	72.1	71.9	**74.9**
Pr→ Ar	38.5	46.1	45.8	56.7	**64.2**	59.7	63.7	61.3
Pr→ Cl	31.2	43.7	43.4	49.2	48.8	**53.1**	51.4	52.6
Pr→ Rw	60.4	68.5	70.3	75.7	79.5	78.4	77.1	**79.9**
Rw→ Ar	53.9	63.2	63.9	68.0	**74.5**	72.4	70.9	73.3
Rw→ Cl	41.2	51.8	52.4	54.0	52.6	**60.0**	57.1	54.2
Rw→ Pr	59.9	76.8	76.8	79.5	82.7	82.9	81.5	**83.2**
Avg	46.1	57.6	58.3	64.9	67.6	67.6	67.3	**68.4**

$-\frac{1}{n_s}\sum_{i=1}^{n_s}\log(\sum_{k=1}^{C} q^*_{s_k}(\sum_{j}^{n_t} a^{st}_{ij}f^t_j)) - \frac{1}{n_s}\sum_{i=1}^{n_s}\log(\sum_{k=1}^{C} q^*_{t_k}(f^s_j))$. Under this circumstance, generator synthesises domain-invariant features by adversarial training. Specifically, we train generator with fixed classifiers by minimizing objective function:

$$\min_{G} \mathcal{L}_s + \mathcal{L}_t + \lambda_1(\mathcal{L}_n + \mathcal{L}_d) + \lambda_2\mathcal{L}_{em}, \tag{2.9}$$

where λ_1 and λ_2 control the relative importance of domain alignment and entropy minimization. Finally, we repeat **Step B**, **Step C** and **Step D** to obtain optimal model.

2.3 Experiment

The proposed method is evaluated on three popular benchmark datasets of unsupervised domain adaptation and compared with other state-of-the-art algorithms.

2.3.1 Experimental Setting

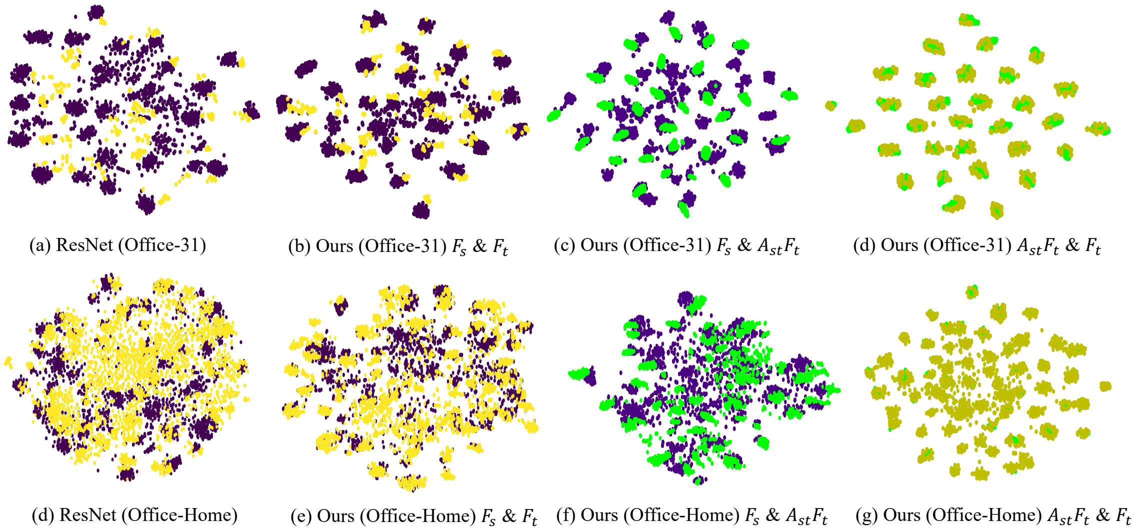
Office-31 is considered as a standard benchmark dataset for UDA problem [85]. It contains 4,110 images collected from three various domains: Amazon Website (**A**), Web camera (**W**) and Digital SLR camera (**D**). Although images of three domains are captured under distinctive conditions, **A**, **W** and **D** share the same label space with 31 categories. In addition, the biggest challenge of domain adaptation in this dataset is imbalanced across three domains. Specifically, Amazon domain consists of 2,817 images, while DSLR domain and Webcam domain only contain 498 and 795 images, respectively. We evaluate six domain adaptation tasks in Office-31.

Office-Home is another more challenging dataset for visual domain adaptation [86]. It includes 15,500 images belonging to 65 categories. These images containing various daily objects are captured in office or home scenes. There are four different domains: Artistic images (Ar), Clip Art (Cl), Product images (Pr) and Real-World images (Rw), which forms 12 adaptation tasks.

ImageCLEF-DA dataset is another popular standard benchmark for unsupervised domain adaptation including three domains: Caltech-256 (**C**), ImageNet ILSVRC 2012 (**I**) and Pascal VOC 2012 (**P**). Arbitrary domain includes 12 categories and each class contains 50 images. Different from Office-Home and Office-31, three domains in this dataset have the same scale. There are six unsupervised domain adaptation tasks to be evaluated.

Baselines: We compare our structure preserving method with generative adversarial algorithms: DANN [36], SymNets [22] and maximum mean discrepancy based on approaches: JAN [27] and other deep models like DSR [84], TADA [82], and SAFN [83]. JAN is implemented with the released code. Moreover, we cite the results of DANN, SymNets, DSR, TADA and SAFN directly from corresponding papers [22, 36, 82, 84] for a fair comparison as we adopt the exact the same experimental protocol.

Implementation details: We implement the proposed method on Tensorflow. The ResNet-50 (without the last FC layer) pre-trained on ImageNet dataset [11] is employed to extract features from raw images. We only fine-tune parameters of ResNet-50 on source domain. The architecture in classifier C_s and C_t both include two-layer FC layers with activation function as *Relu*. We adopt Adam optimizer to update all parameters and select the learning rate $\eta_p = \frac{\eta_0}{(1+ap)^b}$, where p is linearly changing from 0 to 1. We set the initial learning rate $\eta_0 = 0.01$, $\alpha = 10$ and $\beta = 0.75$ according to strategy in [22]. λ_1 and λ_2 are selected from $\{10^{-4}, 10^{-3}, 10^{-2}, 10^{-1}, 1\}$. Finally, we obtain the classification accuracy in target domain using C_t.


(a) ResNet (Office-31) (b) Ours (Office-31) F_s & F_t (c) Ours (Office-31) F_s & $A_{st}F_t$ (d) Ours (Office-31) $A_{st}F_t$ & F_t

(d) ResNet (Office-Home) (e) Ours (Office-Home) F_s & F_t (f) Ours (Office-Home) F_s & $A_{st}F_t$ (g) Ours (Office-Home) $A_{st}F_t$ & F_t

Figure 2.2: Overview of the proposed architecture, where features F_s and F_t are extracted from raw data through generator (VGG or ResNet), and then we capture matching relationship (blue dotted line) of two domains according to graph distribution. Moreover, two classifiers are built and fed with same input. We adopt domain adversarial training manner to maximum the difference between them.

2.3.2 Comparison Results

Table 2.1 shows classification accuracy result of domain adaptation task on Office-31 dataset. The proposed approach overpasses all compared methods in terms of average accuracy. Due to imbalanced condition across three domains, it is difficult for model to transfer knowledge learned in a small-scale dataset into another larger domain. However, different from the results of other algorithms in tasks **D** → **A** and **W** → **A**, our model shows less sensitive to imbalanced circumstance. The main reason for success of our model is that we introduce cross-domain graph information into our method. Alignment with graph discovers similarity of topological structure and utilizes consistency to address domain shift. On the other hand, target classifier with cross-domain graph provides feature learning of target domain with more label information from source domain.

The classification results about 12 domain adaptation tasks on the Office-Home [86] is reported in Table 2.2. As we all know, since office-Home dataset has more categories than office-31 dataset, it is difficult for the same method to produce better result than its performance in office-31 dataset. Compared to ResNet-50 only fine-tuned in source domain, impressive improvements have been obtained with the mentioned methods. The performance of our method significantly achieves improvements when compared with other algorithms. Although the results of SymNets on tasks **Cl** → **Ar**, **Cl** → **Pr** and **Rw** → **Cl** are higher, our method substantially promotes classification accuracy in most cases and obtains better average performance. Specifically, our model produces higher accuracy with large margin for several difficult tasks such as **Ar** → **Cl** and **Ar** → **Pr** task. It indicates that the proposed method effectively eliminates domain discrepancy and extracts domain-invariant feature by graph alignment and domain adversarial alignment.

Table 2.3 reports classification accuracy on ImageCLEF-DA dataset. Different from previous two datasets, each domain in this dataset has the same number of samples. All methods even ResNet-50 totally obtain impressive accuracy. According to comparison with mentioned methods, our model achieves the best performance in most cases e.g., $P \to C$, $C \to I$ and $I \to C$, demonstrating the effectiveness of our proposed method in solving domain adaptation problem. In addition, compared to traditional adversarial training methods (DANN and CDAN), our model and SymNets both perform better results than them, benefiting from symmetric adversarial training manner. Two classifiers in symmetric adversarial method tend to describe the same feature from various perspectives. Thus, the discriminative ability of target classifier is improved dramatically.

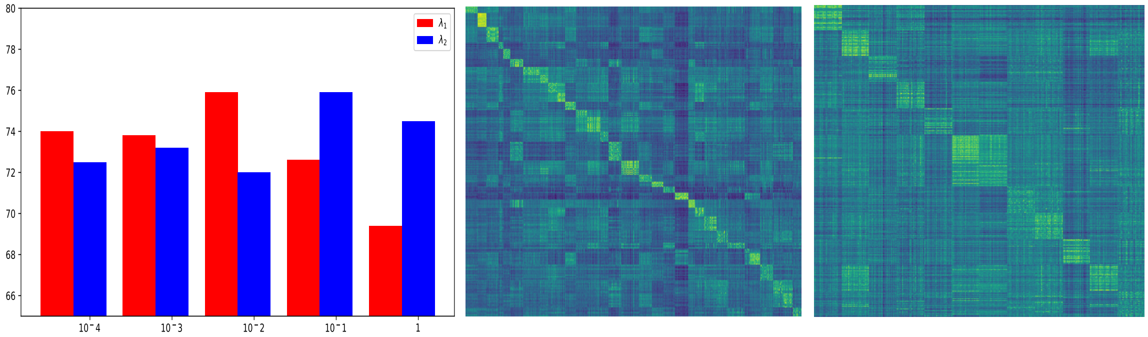
Table 2.3: Top-1 Accuracy (%) on ImageCLEF-DA dataset for UDA (ResNet-50) and the best result is in bold type.

Method	I→ P	P→ I	I→ C	C→ I	C→ P	P→ C
ResNet-50	74.8	83.9	91.5	78	65.5	91.2
DAN	74.5	82.2	92.8	86.3	69.2	89.8
DANN [36]	75	86	96.2	87	74.3	91.5
JAN [27]	76.8	88	94.7	89.5	74.2	91.7
CDAN [38]	76.7	90.6	97	90.5	74.5	93.5
SymNets [22]	**80.2**	93.6	97	93.4	**78.7**	96.4
SAFN [83]	79.3	**93.8**	96.3	91.7	77.6	95.3
Ours	79.4	91.9	**97.9**	**94.1**	76.5	**97.2**

2.3.3 Ablation Study

t-SNE visualization: To understand the effect of graph alignment, we utilize t-SNE visual technique to observe distribution of features in 2D-space. We compute t-SNE with output of the last FC layer in generator and conduct experiments on Office-31 (A→W) and Office-Home (Ar→Cl) for the original ResNet-50 features and our model. According to Fig. 2.2 (a), there are a few overlaps between target instances (yellow) and samples of source domain (purple), demonstrating cross-domain distribution exists large difference named domain shift. Through feature learning phase with GSP, target samples are embedded into source domain in Fig. 2.2 (b). When comparing the location of target samples in Fig. 2.2 (a) and Fig. 2.2 (b), We also know that there is a phenomenon of translation resulting from the influence of graph alignment which matches target samples with source data points. The comparison between F_s and $A_{st}F_t$ is shown in Fig. 2.2 (c). Different from F_t, almost all $A_{st}F_t$ are attached to features of source domain. It illustrates that GSP learns cross-domain matching relation and exploits it to transform target domain into source domain. Since source domain (A) contains more samples than target domain (W), space expanded by $A_{st}F_t$ becomes larger than that of F_t in Fig. 2.2 (d). Thus, reducing domain discrepancy tends to be obstructed with difference between $A_{st}F_t$ and F_t. In addition, focusing on the center of Fig. 2.2 (e), this area are occupied by abundant target samples with a few source instances. GSP employs graph information to discover cross-domain similarity and transfers data points of target domain into the corresponding instances of source domain in Fig. 2.2

(f), meaning our model effectively achieves domain adaptation. Similar with Fig. 2.2 (e), $A_{st}F_t$ mostly are embedded into source domain. The last Fig. 2.2 (h) shows relationship between $A_{st}F_t$ and F_t on office-home dataset. Abundant overlaps between them means they share the same space. Thus, we transform target domain into source domain through $A_{st}F_t$.



(a) Parameter Analysis (b) A_{st} on Office-31 (c) A_{st} on Image-CLEF-DA

Figure 2.3: (a) Parameter analysis of our proposed model GSP. We conduct experiments on Office-31 with task $D \rightarrow A$ and investigate classification accuracy with varying parameters λ_1 and λ_2. (Red: λ_1, Blue: λ_2), (b) Visualization of cross-domain graph A_{st} on Office-31 (D→ W) with 31 categories and (c) Visualization of cross-domain graph A_{st} on ImageCLEF-DA (P→ C) with 12 categories.

Parameter analysis: In this section, we conduct experiments to observe the performance of our model with parameters λ_1 and λ_2. The control variations method is adopted to investigate experimental results. We select value from $\{10^{-4}, 10^{-3}, 10^{-2}, 10^{-1}, 1\}$. Concretely, when fixing parameter λ_1, we change parameter λ_2 from 10^{-4} to 1. The parameter analysis is conducted on Office-31 ($D \rightarrow A$) and Fig. 2.3 (a) reports results. According to Fig. 2.3 (a), as λ_1 goes up, classification accuracy tend to be improved and then be reduced gradually, illustrating our model is sensitive to parameter λ_1 which adjusts importance of domain adversarial term. However, our method becomes stable when raising the value of λ_2. GSP achieves optimal result with $\lambda_1 = 0.01$ and $\lambda_2 = 0.1$.

Cross-domain Graph Analysis: In addition to t-SNE analysis, we also visualize graph matching A_{st} to observe the performance of edge-level alignment which attempts to discover cross-domain matching relation. Ideally, A_{ij}^{st} has large value when f_i^s and f_j^t belong to the same category, otherwise, A_{ij}^{st} tends to be small. We conduct experiments on Office-31 ($\mathbf{W} \rightarrow \mathbf{D}$) and ImageCLEF-DA ($\mathbf{P} \rightarrow \mathbf{C}$) and extract the optimal A_{st} shown in Fig. 2.3 (b) and (c). The visualization of graph exhibits diagonal block structure which means GSP explores edge-level alignment to capture cross-domain matching information.

2.4 Conclusion

In this paper, we rethink Unsupervised Domain Adaptation (UDA) from the perspective of graph distribution and propose Generative Cross-domain learning via Structure Preserving (GSP) to address domain shift problem. GSP model mainly contains two important components: graph alignment and source-supervised target classifier. Graph alignment utilizes edge-level alignment to capture cross-domain matching relation and incorporates relation into node-level alignment to eliminate domain shift. Moreover, we introduce matching information into classifiers and develop source-supervised target classifier exploiting label of source domain to supervise feature learning of target domain. To maximize difference of two classifiers, we adopt symmetric adversarial training manner to train neural network. Extensive experimental results and analyses on several cross-domain visual benchmarks have illustrated the effectiveness of GSP model by comparing with other competitive methods.

Chapter 3

Generative Metric Learning for Unsupervised Domain Adaptation

3.1 Background

Recent years witness the abundant successful applications of deep neural networks (DNN) on computer vision tasks such as image classification [87], video analysis [88] and image semantic segmentation [89]. The achievement undoubtedly attributes to the ability of learning abstract semantic knowledge with the hierarchical network architecture from visual signals [90]. Typically, large-scale datasets like ImageNet [11] become essential for the model training of mainstream frameworks (e.g., ResNet [76] and VGG-Net [91]) with massive learnable parameters. For some specific learning tasks, the model learned from the off-the-shelf well-labeled training set generally suffers from performance degradation when evaluated on the novel test set due to their distribution divergence. The practical dilemma motivates the emerging research topic named as "Unsupervised Domain Adaptation (UDA)" [36].

UDA scenario allows us to train a generalized model with label-sufficient source domain and unlabeled target samples [31, 33]. Its basic assumption is that cross-domain images are collected from different distributions yet share the identical label space [40, 41]. Such a domain shift comes from multiple factors as different illuminations, occlusions or style changes, leading to negative influence on the prediction of target samples [56]. Thus, the

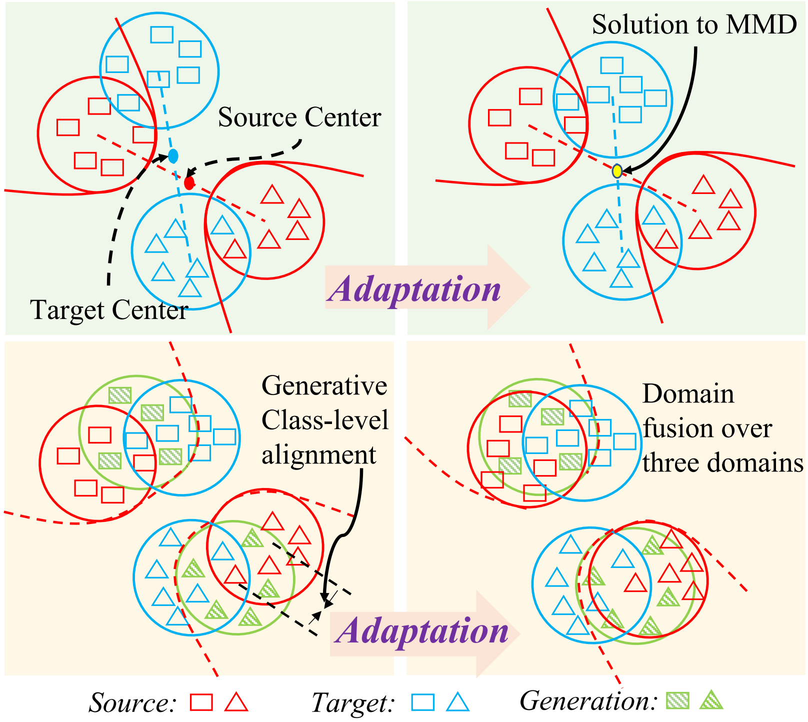


Figure 3.1: Comparison of MMD and MSGD, where the top row shows that MMD calculates the domain-specific center of all observable instances and narrows down their distance, while the bottom row represents that MSGD explores the cross-domain structural knowledge over target samples to explicitly synthesize an intermediate domain with the specific source samples and utilizes it to bridge source and target.

core task of UDA is to borrow sufficient transferable knowledge from source domain to make model generalized well on target domain. To achieve such an expectation, existed approaches typically attempt to minimize the empirical risk with source instances and extract domain-invariant features as cross-domain knowledge [57, 59–61].

Generally, mainstream strategies on solving UDA mainly are categorized into two branches: domain adversarial confusion [22, 64] and metric-based domain alignment [21, 27]. The first branch based on the game theory [92] encourages feature extractor to generate domain-invariant representations to confuse discriminator. However, adversarial mechanism extracts domain-level knowledge yet ignores the improvement of feature discrimination. The other direction focuses on the measurement of domain discrepancy with appropriate metric standards. As shown in upper row of Figure 3.1, DAN [21] deploys the maximum mean discrepancy (MMD) to enforce the consistency of distribution centers across source and target domains. [29] advances MMD by considering multiple explicit match with respect to high order central moments. Similarly, AdaBN [93] incorporates these statistical attributions into batch normalization module to align cross-domain features distributions during forward propagation. Although statistical properties of all observable instances roughly reflect domain shift, it is difficult for them to capture class-level variations due to the absence of target annotations, which negatively affects the learning of robust classifier.

To overcome such a challenge, recent works [33, 94, 95] explore the model prediction and structural information to obtain pseudo labels for target instances during the training stage. This operation creates the predominant condition for class-level alignment. For example, TPNet [94] learns domain-wise prototypes and constructs transferable prototypical module to achieve domain adaptation. Moreover, [30] narrows down intra-class distance and expands inter-class divergence over cross-domain samples to learn discriminative features. In addition, [32] introduces the class prior probability of samples into MMD constraint to handle the imbalanced issue. Although such strategies effectively facilitate class-level distribution alignment, mis-classified pseudo labels tend to trigger inaccurate adaptation in class level.

Beyond class-level alignment, other explorations [34, 96] delve into sample-to-sample distance measurement, which assumes that feature extractor learns similar representations for source and target samples from the same category. Along this line, ETD [34] modifies the optimal transport with attention matrix to capture correlation of cross-domain samples and enhance their representations consistency. However, the sample-level alignment manner difficultly fights off considerable domain mismatch. It is natural to post a question: *"Can we rely on an intermediate domain to gradually eliminate the discrepancy between source and target domains?"*, which has been explored by several literature in early stage. CyCADA [40] based on GAN framework develops the intermediate images by stylizing the source images with target semantic. But the domain shift deriving from other multiple factors (*e.g.*, occlusion, light condition, capturing angle) are neglected during the generative process, which negatively affects the quality of gradual alignment. And DM-ADA [41] randomly mixes up the source and target images to generate the pixel-level intermediate samples. Although DM-ADA can improve the sample diversity with this manner, the generated images are more likely to contain multiple classes information, which tends to confuse the training of classifier.

In this paper, we consider generating the intermediate domain in high-level feature space with the guidance of cross-domain topological structure and treat it as the bridge to narrow down the original domain shift gradually. This is formulated into a novel method named Maximum Structural Generation Divergence (MSGD). Concretely, MSGD first utilizes the output of classifier to capture the intrinsic cross-domain topological knowledge and takes it as the combination coefficient to propagate target samples with the specific source instances. Second, as the bottom of Figure 3.1 shown, the semantic similarity across source and intermediate samples makes the explicit class-level alignment possible to decrease their divergence, while the domain-level measure is favorable for intermediate and target domains. The main contributions are summarized into two folds:

- First of all, we aim to capture the cross-domain topological knowledge based on the discriminative classifier prediction, and deploy it to propagate target samples into the intermediate domain with the guidance of source samples. The intermediate domain is treated as the bridge to gradually reduce distribution divergence across source and target domains.
- Second, we exploit the delicate class-level alignment to eliminate domain shift across source and intermediate domains due to their semantic similarity and domain-level match strategy on intermediate and target domains. To improve the generative quality, we develop the class-driven collaborative translation (CDCT) module to select class-consistent source and target samples in each mini-batch.

3.2 The Proposed Method

3.2.1 Preliminaries and Motivation

Unsupervised Domain Adaptation (UDA) attempts to borrow transferable knowledge from label-sufficient source domain to accurately perform classification on target domain without any annotation. Formally, we have access to a source domain $\mathcal{D}_s = \{(x_i^s, y_i^s)\}_{i=1}^{n_s}$ including n_s samples with the corresponding labels as well as an unlabeled target domain $\mathcal{D}_t = \{x_i^t\}_{i=1}^{n_t}$

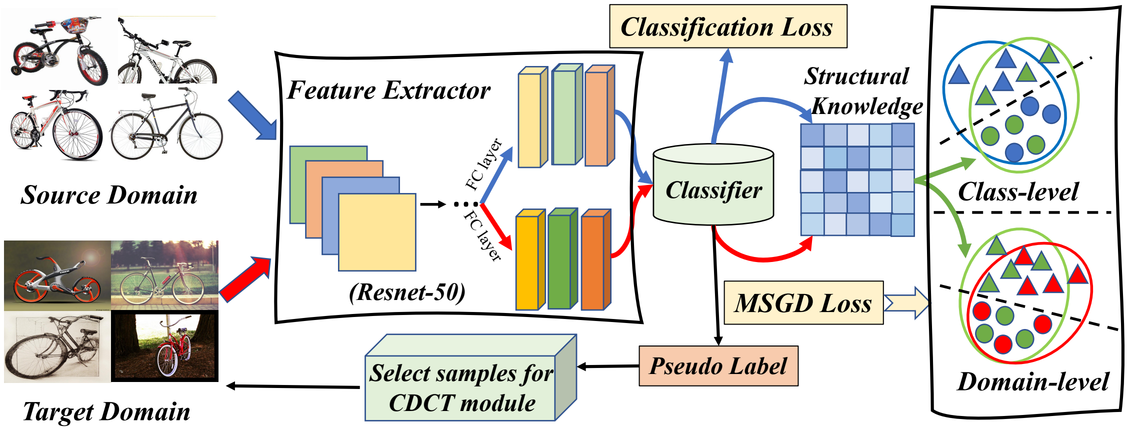


Figure 3.2: Overview of the proposed architecture. The feature extractor aims to learn domain-invariant representations as the input of classifier. The output of classifier needs to achieve three tasks. First, we exploit the source output to calculate the classification loss. Second, the target prediction is used as pseudo label to select samples for Class-Driven Collaborative Translation (CDCT) module. Moreover, we explore all outputs from various domains to construct structural knowledge and combine it with target features to build the intermediate domain (Green). Finally, we separately adopt class-level and domain-level alignments to conduct domain fusion between intermediate domain and source or target domain.

with n_t instances. The general assumption of UDA is that different domains belong to non-iid distributions, i.e., $x^s \sim P_s$ and $x^t \sim P_t$, but they share the identical label space. Such a domain shift results in the performance degradation of learned source-supervised model on the target recognition.

The intuition to align different distributions is adjusting cross-domain feature representations to reach the similar even identical statistics properties. Along with this direction, MMD [97] measures and eliminates the cross-domain divergence by comparing their domain-wise mean in the reproducing kernel hilbert space (RKHS). In addition, generative adversarial strategy comprehensively captures the attributions of distribution and approximates it instead of depending on specific pre-defined indicators [38, 39]. Actually, these existing approaches effectively promote the model generalization by reducing domain-level discrepancy and learning more transferable features. However, it is difficult for such an alignment manner to obtain consistent or similar representations for cross-domain samples potentially from the same category, especially for the considerable distribution shift between source and target domains.

Towards the practical bottleneck, we naturally post a question ***"Can we construct an***

intermediate domain between source and target domains to gradually achieve domain alignment?" To answer this, we have to grasp the core properties of the intermediate domain. That is, the intermediate domain needs to bridge the gap between source and target domains by narrowing down their distribution difference. In fact, there are several literature start thinking about the similar question in early stage like generative model [40] and mixup strategy [41]. Differently, we consider generating the intermediate instances by exploring the cross-domain topological structure. Specifically, we expect to propagate category-wise target samples to generate the augmented instances naturally paired with per given source sample. The similar semantic information across source and intermediate instances makes the class-level alignment straightforward, while the domain-level matching is proper for eliminating the distance between intermediate and target domains due to the absence of ground-truth. These considerations are formulated as our Maximum Structural Generation Discrepancy (MSGD).

3.2.2 Maximum Structural Generation Divergence

Per discussion above, the key becomes how to actually capture the cross-domain geometric knowledge to synthesize novel class-specific source-like samples conditioned on target ones. However, without target annotations, it is difficult for us to select the target instances from the same category with source samples to conduct the imitation. But we also notice that the ideal instance integration is to endure large combination coefficient for two within-class cross-domain samples. On the other hand, the semantic-similar source and target samples are closer to each other on the topological structure and have the larger sample-wise similarities. With these observations, we naturally consider such intrinsic structural knowledge as the coefficients to explore the intermediate domain.

Compared with the raw inputs, the high-level deep features are promising to reflect the sample-wise association in the compressed low-dimensional space. To benefit from more discriminative information, we aim to explore the cross-domain topological knowledge from the high-level classifier prediction. Concretely, given source and target features within a mini-batch $\mathbf{F}^s = [f(x_1^s), f(x_2^s), ..., f(x_{b_s}^s)]^\intercal$ and $\mathbf{F}^t = [f(x_1^t), f(x_2^t), ..., f(x_{b_t}^t)]^\intercal$ where $f(x_i^{s/t}) \in \mathbb{R}^d$ means the high-level features via the mapping function $f(\cdot)$ and $b_s(b_t)$ is the batch size. The classifier of our framework in Figure 3.2 takes them as input to assign class label probabilities as $\mathbf{Q}^s = [q(x_1^s), q(x_2^s), ..., q(x_{b_s}^s)]^\mathsf{T}$ and $\mathbf{Q}^t = [q(x_1^t), q(x_2^t), ..., q(x_{b_t}^t)]^\mathsf{T}$, where $q(x_i^{s/t}) \in \mathbb{R}^C$ and C is the number of category. Therefore, the cross-domain structural knowledge is defined as $\mathbf{A}$ originating from the normalized cosine similarity:

$$\mathbf{A}_{ij} = \frac{\langle q(x_i^s), q(x_j^t) \rangle}{\|q(x_i^s)\|_2 \|q(x_i^t)\|_2}, \tag{3.1}$$

where $\mathbf{A}_{ij}$ with large value indicates x_i^s and x_j^t are with high probability from the same category, showing the cross-domain intrinsic structure. To ensure the consistence of feature scale, each row of $\mathbf{A}$ is further normalized into $\|\mathbf{A}_i\|_2 = 1$. With $\mathbf{A}$ as translation coefficient, the target-to-source generative operation is formulated as:

$$\mathbf{F}^g = \mathbf{A}\mathbf{F}^t, \quad \mathbf{F}^g = [f_1^g, f_2^g, ..., f_{b_s}^g]^\mathsf{T}, \tag{3.2}$$

where $f_i^g \in \mathbb{R}^d$ and d is the feature dimension. The construction of intermediate domain is similar with the embedding propagation [98], which obtains the combination coefficient by conducting label propagation [99] on the Laplacian of the adjacency matrix calculated from the high-level features $f(x_i^{s/t})$. Compared with embedding propagation, our method learns the cross-domain structural knowledge from the classifier output including more discriminative information than the hidden features, which easily captures the sample-wise intrinsic relation to produce the high-quality instances. To this end, we achieve two important observations. First, structural generative instances constitute a novel domain with distribution being different from that of source or target domain, i.e., $f_i^g \sim P_g$. Second, although the representation of f_i^g is not exactly the same with that of f_i^s, they are more likely to contain the similar even identical category attribution due to the guidance of structural knowledge.

According to the above analysis, the class-wise mean discrepancy is suitable to measure their domain difference when compared with domain-level alignment, so that we design the following formulation:

$$\begin{aligned}\mathcal{D}^2_{\mathcal{H}}(P_s, P_g) \triangleq \sup_{f\sim\mathcal{H}} \Big(&\|\mathbb{E}_{x^s\sim P_s^{c1}}(f^s) - \mathbb{E}_{f^g\sim P_g^{c1}}(f^g)\|^2_{\mathcal{H}} \\ &+ \cdots + \|\mathbb{E}_{x^s\sim P_s^{cn}}(f^s) - \mathbb{E}_{f^g\sim P_g^{cn}}(f^g)\|^2_{\mathcal{H}}\Big),\end{aligned} \tag{3.3}$$

where $P^{c_i}_{s/g}$ denotes the distribution of the i-th category in source or intermediate domain, $\mathcal{H}$ means the RKHS and $f^s = f(x^s)$. To clearly reformulate it into empirical risk loss, we firstly define the class-wise indicator matrix $\mathbf{M} \in \mathbb{R}^{n_s\times n_s}$, where $m_{ij} = 1$ when $y^s_i = y^s_j$, otherwise, $m_{ij} = 0$. The kernel matrices related to intra-domain and inter-domain are calculated by:

$$\begin{aligned}\widetilde{\mathbf{K}}(F^{s/g}, F^{s/g}) &= \mathbf{M} \odot \mathbf{K}(F^{s/g}, F^{s/g}), \\ \widetilde{\mathbf{K}}(F^{s}, F^{g}) &= \mathbf{M} \odot \mathbf{K}(F^{s}, F^{g}),\end{aligned} \tag{3.4}$$

where $\mathbf{K}_{ij} = \langle f_i, f_j\rangle$ and $\odot$ denotes the element-wise product. The final formulation of the empirical estimator with kernel mean embedding is:

$$\begin{aligned}\mathcal{D}^2_{k}(P_s, P_g) =& \frac{1}{b_s^2}\sum_{i,j=1}^{bs} \Big(\widetilde{\mathbf{K}}(F^s, F^s) + \\ &\widetilde{\mathbf{K}}(F^g, F^g) - 2\widetilde{\mathbf{K}}(F^s, F^g)\Big).\end{aligned} \tag{3.5}$$

So far, Eq. (3.5) mitigates the domain mismatch across source and intermediate domains, and we further align the target and intermediate domains to eventually reduce domain shift between source and target domains. Due to the absence of target annotation, the class-level adaptation becomes invalid. Alternatively, we attempt to measure their distance via domain-wise discrepancy:

$$\begin{aligned}\mathcal{D}^2_{k}(P_t, P_g) =& \frac{1}{b_t^2}\sum_{i,j=1}^{b_t} \mathbf{K}(F^t, F^t) + \frac{1}{b_s^2}\sum_{i,j=1}^{b_s} \mathbf{K}(F^g, F^g) \\ &- \frac{2}{b_t\cdot b_s}\sum_{i,j=1}^{b_t, b_s} \mathbf{K}(F^t, F^g).\end{aligned} \tag{3.6}$$

To this end, we have formulated the direct source-to-target domain discrepancy [96] into two finer parts to measure the difference of source-to-intermediate and intermediate-

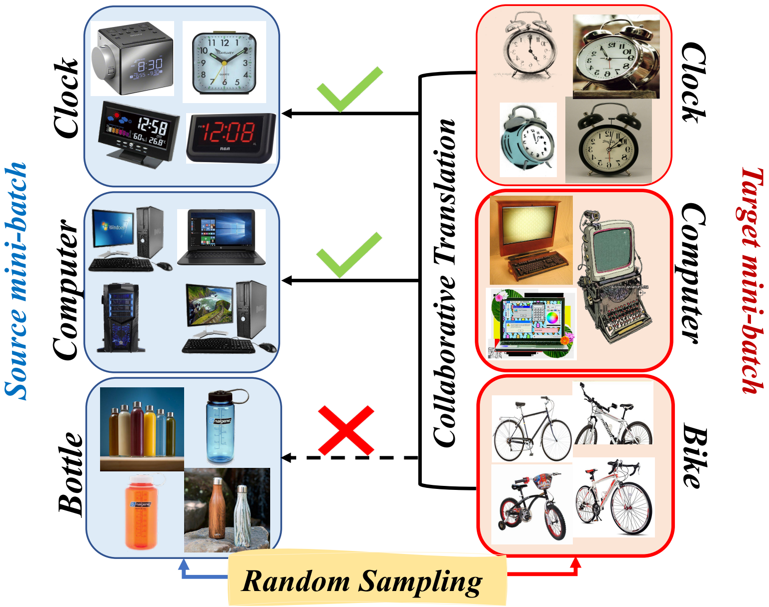


Figure 3.3: Problem of traditional sampling manner on collaborative representation. Concretely, the category "Bottle" is randomly selected in source mini-batch but does not exist in target mini-batch. Generative instances via collaborative translation from target mini-batch are corresponding to "Bottle" but tend to be far from this category.

to-target in a more flexible way. Hence, we propose a unified framework named Maximum Structural Generation Divergence (MSGD) as:

$$\mathcal{L}_d = (1-\alpha)\mathcal{D}_k^2(P_s, P_g) + \alpha\mathcal{D}_k^2(P_t, P_g), \tag{3.7}$$

where $\alpha \in (0,1)$ is a trade-off parameter to balance triplet domain alignment.

Remarks: The strict node-level alignment in [96] is adopted to eliminate sample-to-sample representation difference. Such a constraint clusters source and target samples from the same category into the same subspace. However, it tends to destroy the diversity of generative sample and have a negative influence on the generalization of model. The proposed model effectively addresses this issue by comparing the distributions of source and intermediate domains.

Class-Driven Collaborative Translation

So far, we propose the novel Maximum Structural Generation Divergence to mitigate the domain shift with the guidance of intermediate domain. To achieve effective alignment, the key is to build a discriminative graph $\mathbf{A}$. In the implementation of DNN, mini-batch strategy is usually adopted to handle large-scale dataset. Hence, within each mini-batch, the number of categories selected from source domain might be less or more than that in target domain, or source and target batches involve the same class number yet various categories. For example, the combination of samples in target batch is exploited to represent source sample whose class disappears in $\mathbf{F}^t$ as Figure 3.3. These situations provide wrong guidance for the target-to-source collaborative translation ($\mathbf{F}^g = \mathbf{A}\mathbf{F}^t$), which motivates us to propose Class-Driven Collaborative Translation (CDCT).

Specifically, CDCT involves two primary tasks: 1) to annotate target samples with structural information and 2) to filter uncertain target samples according to the category in $\mathbf{F}^s$. For every epoch, all source and target samples are fed into the current model to generate hidden features $\{f(x_i^{s/t})|i = 1, 2, ..., n_{s/t}\}$. According to the guidance of ground truth, it is simple to calculate source class center $\mathcal{C}_j^s = \sum_{i=1}^{n_s} \mathbf{I}(y_i^s = c_j) f(x_i^s)/|c_j|$, where $\mathbf{I}(\cdot)$ is the indicator and $|c_j|$ means the number of samples in the j-th category. Since our proposed collaborative translation aims to convert target samples towards source domain, we directly exploit $\mathcal{C}_j^s$ as target class center $\mathcal{C}_j^t$ to measure the sample-to-center distance with:

$$d(f(x_i^t), \mathcal{C}_j^t) = \frac{f(x_i^t)^\top \mathcal{C}_j^t}{\|f(x_i^t)\|_2 \|\mathcal{C}_j^t\|_2}. \tag{3.8}$$

Eventually, we formulate the class probability distribution $p_i^t \in \mathbb{R}^C$ of $f(x_i^t)$ and access to the pseudo labels by:

$$p_{i,j}^t = \frac{\exp(d(f(x_i^t), \mathcal{C}_j^t))}{\sum_{j=1}^{c_n} \exp(d(f(x_i^t), \mathcal{C}_j^t))}, \quad y_i^t = \max_j \{p_{i,j}^t\}. \tag{3.9}$$

According to categories selected in source mini-batch, we use pseudo label to choose target samples from the same categories. However, due the domain shift, unreliable structural knowledge mistakenly labels several target samples. The intuitive solution for this problem is to abnegate target instances with large uncertainty within each class. Thus, we only accept samples whose class probability maximum is larger than a threshold, i.e., $\{x_i^t| \max \{p_{i,j}^t\} \geq \delta\}$ with $\delta = 0.2$ for all experiments. And various categories are corre-

Algorithm 1 Training and Inference of MSGD

```
Input: $\mathcal{D}_s = \{(x_i^s, y_i^s)\}_{i=1}^{n_s}$, $\mathcal{D}_t = \{x_i^t\}_{i=1}^{n_t}$.
Initialization: $\theta, \alpha$.
Pre-train the network with source data via classification loss
For epoch IN range(K):
    Calculate source centers and annotate target samples;
    For iteration IN range(T):
        Randomly select source sample for mini-batch;
        Select target sample via CDCT module;
        Compute source-to-generative discrepancy via Eq. (3.5);
        Compute target-to-generative discrepancy via Eq. (3.6);
        Update the network via Eq. (3.10);
    End
End
Output: Model Parameters and Target Prediction.
```

sponding to different threshold. In practice, we collect the class probability maximum of each instance and regard their average as the threshold. Benefiting from these operations, we successfully conduct class-driven collaborative translation to assist maximizing structural generation divergence.

3.2.3 Overall Objective

The previous triplet domain alignment module mainly explores cross-domain structural knowledge to construct intermediate domain and applies the novel domain to eliminate source-to-target domain discrepancy, which is helpful for feature extractor to learn domain-invariant representation. And then, the neural network classifier needs to identify the category of these features. Similar with the existed works [38, 96], we integrate the source supervision to train a classifier and formulate our overall objective as:

$$\min_{\theta} \ \mathcal{L} = \frac{1}{n_s} \sum_{i=1}^{n_s} \mathcal{L}_c\big(f(x_i^s), y_i^s\big) + \mathcal{L}_d, \tag{3.10}$$

where θ denotes the set of all the parameters to be optimized in feature extractor and classifier, $\mathcal{L}_c$ denotes the source classifier. Finally, the details of model training with our MSGD are summarized in **Algorithm 1**.

Table 3.1: Classification Accuracy (%) on Office-31 for unsupervised domain adaptation tasks (Network backbone: Resnet-50). The best result among all competitive methods is highlighted with **bold** type, while the second performance is marked with <u>underline</u>.

Office-31	A→ W	D→ W	W→ D	A→ D	D→ A	W→ A	Avg
Resnet-50 [76]	68.4±0.2	96.7±0.1	99.3±0.1	68.9±0.2	62.5±0.3	60.7±0.3	76.1
CDAN [38]	94.1±0.1	98.6±0.1	**100.0±0.0**	92.9±0.2	71.0±0.3	69.3±0.3	87.7
SAFN [83]	88.8±0.4	98.4±0.0	**99.8±0.0**	87.7±1.3	69.8±0.4	69.7±0.2	85.7
Symnet [22]	90.8±0.1	98.8±0.3	**100.0±0.0**	93.9±0.5	74.6±0.6	72.5±0.5	88.4
ALDA [100]	**95.6±0.5**	97.7±0.1	**100.0±0.0**	94.0±0.4	72.2±0.4	72.5±0.2	88.7
DADA [39]	92.3±0.1	<u>99.2±0.1</u>	**100.0±0.0**	93.9±0.2	74.4±0.1	74.2±0.1	89.0
DMP [101]	93.0±0.3	99.0±0.1	**100.0±0.0**	91.0±0.4	71.4±0.2	70.2±0.2	87.4
GVB-GD [102]	94.8±0.5	98.7±0.3	**100.0±0.0**	<u>95.0±0.4</u>	73.4±0.3	73.7±0.4	89.3
DMRL [103]	90.8±0.3	99.0±0.2	**100.0±0.0**	93.4±0.5	73.0±0.3	71.2±0.3	87.9
ETD [34]	92.1	**100.0**	**100.0**	88.0	71.0	67.8	86.2
SE-CC [33]	90.7	99.0	**100.0**	91.4	74.0	72.9	88.0
GSP(Ours) [96]	92.9	98.7	<u>99.8</u>	94.5	<u>75.9</u>	<u>74.9</u>	<u>89.5</u>
MSGD (Ours)	<u>95.5±0.5</u>	<u>99.2±0.3</u>	**100.0±0.0**	**95.6±0.3**	**77.3±0.4**	**77.0**±0.5	**90.8**
Oracle	99.7	99.7	100.0	100.0	85.1	85.1	-

3.3 Experiment

3.3.1 Benchmark Datasets

In experiments, we evaluate our MSGD on five cross-domain learning benchmarks. Office-31, Office-Home and Image-CLEF have been introduced in Chapter 2. The description of VisDA-2017 and DomainNet are described as follows.

VisDA-2017 [104] is a large-scale benchmark dataset for domain adaptation task. The challenge of VisDA is to borrow transferable knowledge from synthetic image data to real-world scenes. It totally includes three sets (training, testing and validation sets) with 280,157 images divided into 12 categories such as train, truck and motorcycle. The 152,397 synthetic visual signals generated from 3D model make up the training set, while the validation set collects 55,388 real images from Microsoft COCO database [105].

DomainNet is currently the largest benchmark dataset for domain adaptation with 590K images from 345 categories and contains six domains as Clipart (**clp**), Infograph (**inf**), Painting (**pnt**), Quickdraw (**qdr**), Real (**rel**) and Sketch (**skt**). Each domain includes training and test sets without overlap. Thus, we follow the protocol of [106] to carry out 30

Table 3.2: Classification Accuracy (%) on Office-Home for unsupervised domain adaptation tasks (Network backbone: Resnet-50). The best result among all competitive methods is highlighted with **bold** type, while the second performance is marked with underline.

Source	Art (Ar)			Clipart (Cl)			Product (Pr)			Real-World (Rw)			Avg
Target	Cl	Pr	Rw	Ar	Pr	Rw	Ar	Cl	Rw	Ar	Cl	Pr	
Resnet-50 [76]	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
CDAN [38]	50.7	70.6	76.0	57.6	70.0	70.0	57.4	50.9	77.3	70.9	56.7	81.6	65.8
SAFN [83]	52.0	71.7	76.3	64.2	69.9	71.9	63.7	51.4	77.1	70.9	57.1	81.5	67.3
Symnet [22]	47.7	72.9	78.5	64.2	71.3	74.2	64.2	48.8	79.5	74.5	52.6	82.7	67.6
ALDA [100]	53.7	70.1	76.4	60.2	72.6	71.5	56.8	51.9	77.1	70.2	56.3	82.1	66.6
DMP [101]	52.3	73.0	77.3	64.3	72.0	71.8	63.6	52.7	78.5	72.0	57.7	81.6	68.1
GVB-GD [102]	57.0	74.7	79.8	64.6	74.1	74.6	65.2	55.1	81.0	74.6	59.7	84.3	70.4
ETD [34]	51.3	71.9	**85.7**	57.6	69.2	73.7	57.8	51.2	79.3	70.2	57.5	82.1	67.3
GSP (Ours) [96]	56.8	75.5	78.9	61.3	69.4	74.9	61.3	52.6	79.9	73.3	54.2	83.2	68.4
MSGD (Ours)	**58.7**	**76.9**	78.9	**70.1**	**76.2**	**76.6**	**69.0**	**57.2**	**82.3**	**74.9**	**62.7**	**84.5**	**72.4**
Oracle	75.0	90.1	87.0	78.7	90.1	87.0	78.7	75.0	87.0	78.7	75.0	90.1	-

Table 3.3: Classification Accuracy (%) on Image-CLEF for unsupervised domain adaptation tasks (Network backbone: Resnet-50). The best result among all competitive methods is highlighted with **bold** type, while the second performance is marked with underline.

Image-CLEF	I→ P	P→ I	I→ C	C→ I	C→ P	P→ C	Avg
Resnet-50 [76]	74.8±0.3	83.9±0.1	91.5±0.3	78.0±0.2	65.5±0.3	91.2±0.3	80.7
JAN [21]	76.8±0.4	88.0±0.2	94.7±0.2	89.5±0.3	74.2±0.3	91.7±0.3	85.8
CDAN [38]	77.7±0.3	90.7±0.2	97.7±0.3	91.3±0.3	74.2±0.2	94.3±0.3	87.7
SAFN [83]	78.0±0.4	91.7±0.5	96.2±0.1	91.1±0.3	77.0±0.5	94.7±0.3	88.1
Symnet [22]	80.2±0.3	93.6±0.2	97.0±0.3	93.4±0.3	78.7±0.3	96.4±0.1	89.9
DMP [101]	80.7±0.1	92.5±0.1	97.2±0.1	90.5±0.1	77.7±0.2	96.2±0.2	89.1
DMRL [103]	77.3±0.4	90.7±0.3	97.4±0.3	91.8±0.3	76.0±0.5	94.8±0.3	88.0
ETD [34]	**81.0**	91.7	97.9	93.3	79.5	95.0	89.1
CAN+A2LP [31]	79.8	94.3	97.7	93.0	**79.9**	96.9	90.3
GSP (Ours) [96]	79.4	91.9	97.9	94.1	76.5	97.2	89.5
MSGD (Ours)	80.2±0.2	**95.7±0.6**	**98.0±0.3**	**94.2±0.3**	79.3±0.3	**97.7±0.2**	**90.9**
Oracle	84.0	97.4	99.4	97.4	84.0	99.4	-

domain adaptation tasks. Specifically, with one domain as target, we respectively consider one of the left five domains as source, and only report the best performance for the specific target domain to make explicit comparison among these methods.

3.3.2 Experimental Setup

Implementation details: As Figure 3.2 shows, we consider Resnet-50 or Resnet-101 pre-trained on ImageNet as the backbone to extract convolutional representations followed by a single fully-connected (FC) layer. Concretely, we carry out experiments on VisDA-2017 dataset with Resnet-101 framework and on other datasets with Resnet-50 architecture. Finally, the classifier exploits another FC operation to generate the final prediction. For the training stage, we adopt stochastic gradient descent (SGD) with momentum of 0.9 as optimizer to update network parameters and set the batch size to be 30. In addition, the annealing strategy [36] is explored to adjust the learning rate by $\eta_p = \frac{\eta_0}{(1+\Lambda p)^\beta}$, where p linearly changes from 0 to 1 according to the progress of training, $\eta_0 = 0.01$, $\Lambda = 10$ and $\beta = 0.75$. We implement experiments on Pytorch platform with one GPU (NVIDIA Titan V). For the hyper-parameter (α) selection, we first introduce a binary domain classifier to distinguish source samples from target ones and follow [21] to jointly assess the test errors of source classifier and domain classifier to determine the optimal α. Concretely, given the specific α, we utilize 90% source instances and all target ones to train the model and evaluate the well-trained model on the left 10% source samples to obtain the test error of source classifier. Moreover, we feed the trained source and target samples into the final model to obtain their high-level features, on which the domain classifier depends to calculate the $\mathcal{A}$-distance with the test error. Finally, we select the optimal α with the minimal sum of source classifier error and $\mathcal{A}$-distance.

Competitive methods: To verify the effectiveness of our method, we compare the performance of MSGD with state-of-the-art algorithms. Specifically, the representative works based on generative adversarial are DANN [36], ADDA [107], CDAN [38], Symnet [22], ALDA [100] and DADA [39]. In terms of metric measurement, we select JAN [27], MCD [70], ETD [34], DMP [101] and GSP [96] as baselines. Moreover, the competitive algorithms also involve other techniques of domain alignment such as DCAN [108], GDCAN [106],

Table 3.4: Classification Accuracy (%) on VisDA-2017 for unsupervised domain adaptation tasks (Network backbone: Resnet-101). The best result among all competitive methods is highlighted with **bold** type, while the second performance is marked with underline.

VisDA-2017	plane	bicycle	bus	car	horse	knife	mcycle	person	plant	sktbrd	train	truck	Avg
Resnet-101 [76]	55.1	53.3	61.9	59.1	80.6	17.9	79.7	31.2	81.0	26.5	73.5	8.5	52.4
JAN [27]	75.7	18.7	82.3	**86.3**	70.2	56.9	80.5	53.8	92.5	33.2	84.5	**54.5**	65.7
CDAN [38]	85.2	66.9	83.0	50.8	84.2	74.9	88.1	74.5	83.4	76.0	81.9	38.0	73.7
SAFN [83]	93.6	61.3	84.1	70.6	94.1	79.0	**91.8**	79.6	89.9	55.6	**89.0**	24.4	76.1
ALDA [100]	93.8	74.1	82.4	69.4	90.6	87.2	89.0	67.6	93.4	76.1	87.7	22.2	77.8
DMP [101]	92.1	75.0	78.9	75.5	91.2	81.9	89.0	77.2	93.3	77.4	84.8	35.1	79.3
SE [109]	96.2	**87.8**	**84.4**	66.5	**96.1**	96.1	90.5	81.5	95.3	91.5	87.5	51.6	85.4
SE-CC [33]	96.3	86.5	82.4	81.3	**96.1**	**97.2**	91.2	**84.7**	94.4	94.1	88.3	53.4	**87.2**
MSGD (Ours)	**97.5**	83.4	**84.4**	69.4	95.9	94.1	90.9	75.5	**95.5**	**94.6**	88.1	44.9	84.6

CAN+A2LP [31], SE-CC [33], SAFN [83], GVB-GD [102] and DMRL [103]. For the fair comparison, we directly report experimental results of baselines according to their published paper as we exactly use the same experimental protocol.

3.3.3 Comparison Results

The object recognition results on Office-31, Office-Home, Image-CLEF, VisDA-2017 and DomainNet are separately summarized in Table 3.1, Table 3.2, Table 3.3, Table 3.4 and Table 3.5. With respect to the average accuracy, the proposed MSGD achieves significant improvement over other state-of-the-art algorithms in most benchmarks. It illustrates that our method effectively mitigates cross-domain discrepancy to improve the generalization of model on target domain. For the following, we delve into specific tasks to achieve more delicate conclusions.

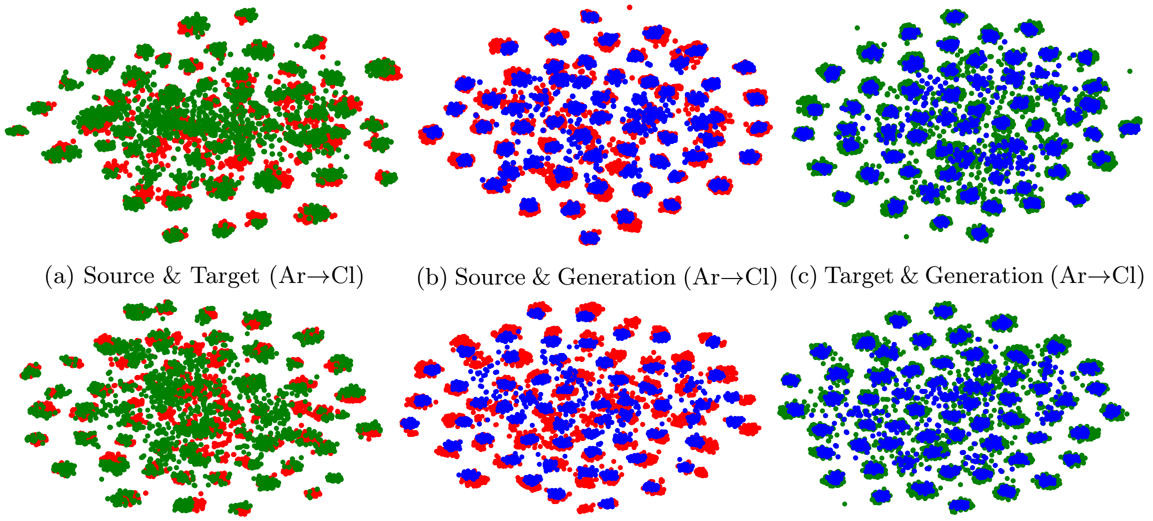
The aforementioned discussion of Office-31 demonstrates there exists imbalanced challenge across various domains *e.g.,* D→A and W→A tasks. Due to insufficient source instances, the model fails to capture the real marginal distribution which has a negative influence on the alignment of joint distribution. Thus, most competitive methods (CDAN, SAFN, ETD) difficultly overcome such problem to obtain better domain adaptation. However, our MSGD substantially promotes the classification accuracy to **77.3**% and **77.0**% for D→A and W→A tasks, respectively. The achievement of MSGD mainly results from the

Table 3.5: Classification Accuracy (%) on DomainNet for unsupervised domain adaptation tasks (Network backbone: Resnet-101). The best result among all competitive methods is highlighted with **bold** type, while the second performance is marked with underline.

DomainNet	clp	info	pnt	qdr	rel	skt	Avg
Resnet-101 [76]	48.4	22.2	49.4	11.1	54.5	38.8	37.4
MCD [70]	42.6	19.6	42.6	3.8	50.5	33.8	32.2
DANN [36]	42.4	16.4	43.1	12.3	48.4	30.4	32.2
ADDA [107]	39.5	14.5	29.1	12.1	41.9	30.7	28.0
DCAN [108]	57.6	19.7	50.5	17.1	60.3	45.8	41.8
GDCAN [106]	58.3	21.8	50.7	17.7	60.8	46.2	42.6
MSGD (Ours)	**60.2**	**23.8**	**53.3**	**20.7**	**62.4**	**48.5**	**44.8**

generation of intermediate domain. Benefiting from the structural knowledge, the synthesised instances are close to the corresponding source samples, which dramatically enhances the diversity of sample. The augmented samples facilitate the model to easily observe the real distribution difference. In addition, we notice that our conference version GSP also explores generative strategy to improve generalization of classifier. But the sample-to-sample constraint is too strict to decrease the generative effect when compared with MSGD. And our model not only significantly fights off the recent work SE-CC [33] on average classification accuracy but also surpasses it by 4.1% on task W→A.

In terms of the results of Office-Home, two conclusions are summarized. Firstly, we have the observation that all methods suffer from the larger performance degradation for the recognition of target sample than that on Office-31 dataset. The main reasons for such situation primarily come from two folds: a) Office-Home involves more categories than Office-31; b) visual signals across various domains are very dissimilar, which triggers the difficulty of learning explicit classification boundary. Secondly, it is worth that our MSGD significantly improves the classification accuracy on most adaptation tasks *e.g.*, Pr→Cl and Rw→Cl under such barren condition. The advantages of MSGD over others are summarized in two points. MSGD synthesizes novel instances similar to source sample and focuses on the alignment of joint distributions about feature and class across source and intermediate domains, which eliminates intra-class distance to form compact category space. In addition, MSGD exploits intrinsic structural information to connect source and target domains by achieving the distribution consistency between target and intermediate domains.



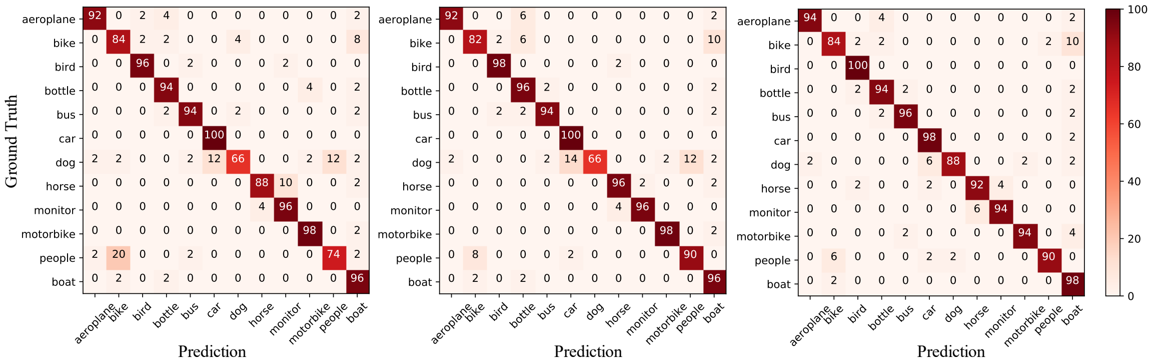
(a) Source & Target (Ar→Cl) (b) Source & Generation (Ar→Cl) (c) Target & Generation (Ar→Cl)

(d) Source & Target (Pr→Cl) (e) Source & Generation (Pr→Cl) (f) Target & Generation (Pr→Cl)

Figure 3.4: Feature visualization of the training process on Office-Home dataset. (**a**): T-SNE of Source and Target domains (Ar→Cl). (**b**): T-SNE of Source and Intermediate domains (Ar→Cl). (**c**): T-SNE of Target and Intermediate domains (Ar→Cl). (**d**): T-SNE of Source and Target domains (Pr→Cl). (**e**): T-SNE of Source and Intermediate domains (Pr→Cl). (**f**): T-SNE of Target and Intermediate domains (Pr→Cl). T-SNE is calculated with the output of feature extractor. Red, Blue and Green indicate source, intermediate and target domains, respectively.

From Table 3.3, although JAN and MSGD both explore maximum mean discrepancy to learn domain-variant feature representation, the classification accuracy of MSGD surpasses that of JAN by a large margin such as the improvements on tasks P→I (**8.7**%) and C→P (**5.1**%). It highly affirms the efficiency of MSGD on accurately estimating the discrepancy of various distributions through the generation of novel sample. Different from GSP learning cross-domain relation to achieve sample-level alignment, our MSGD introduces an intermediate status between source and target domains and attempts to enforce them into such common situation. MSGD thus avoids sample-to-sample mismatch to better reduce domain shift and facilitates the generalization of model on target domain. Moreover, even though CAN+A2LP [31] utilizes label propagation manner to promote model performance, our MSGD still achieves more promising results over it, which illustrates our method effectively transfers more knowledge from source domain to identify target images.

Table 3.4 reports the experimental performances of our method and other baselines on VisDA-2017 benchmark dataset. For the average accuracy, our method fights off DRMEA with a large margin as **5.3**%, which illustrates that MSGD effectively solves domain adap-



(a) Confusion Matrix (MMD)

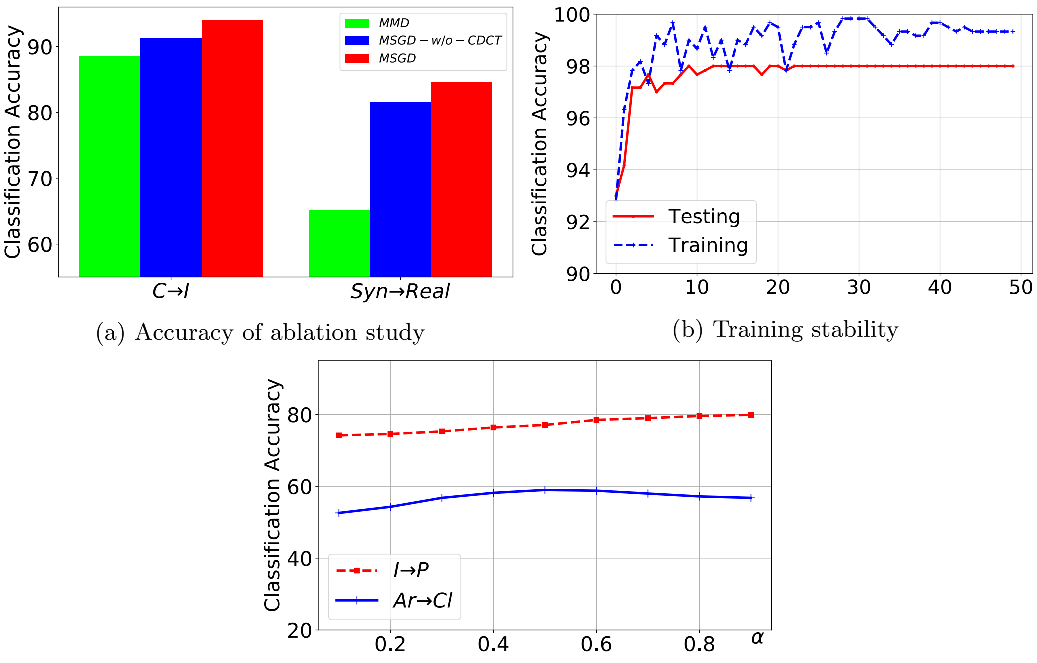
(b) Confusion Matrix (MSGD w/o CDCT)

(c) Confusion Matrix (MSGD)

Figure 3.5: Confusion Matrix reporting the accuracy (%) of the prediction and ground truth. Experiments are performed on Image-CLEF dataset about task C→I. (**a**): Replace our MSGD with MMD directly constraining source and target domains. (**b**): Remove CDCT from MSGD. (**c**): Our proposed MSGD.

tation task on large-scale datasets. Compared to JAN, our MSGD achieves significant improvement of classification accuracy on most categories. Specifically, MSGD promotes the accuracy from 18.7% to 83.4% for class “bicycle”. Even though the average classification accuracy of our MSGD is lower than that of SE and SE-CC, our MSGD still achieves the comparable performance with them in most categories, especially for category ”bus” and ”plane”, our MSGD obtains the best result over others. Compared with them, our MSGD mainly fails to effectively transfer the knowledge of ”person” and ”truck”. Interestingly, JAN [27] as a classical UDA method achieves good performance on ”car” and ”truck”, which means different models have disadvantages and advantages in different categories.

With respect to the experimental results in Table 3.5, although DomainNet includes abundant complicated scenarios and more categories, our MSGD still achieves better performance than other competitors on the average classification accuracy. Specifically, our method outperforms the second best result obtained by GDCAN by 2.8%. In addition, we notice that it is difficult for these methods to achieve promising cross-domain adaptation from other domains to Quickdraw (**qdr**). However, for the challenging tasks, our MSGD facilitates model to obtain higher classification accuracy than baselines. Moreover, for the most challenging task Real (**rel**) to Quickdraw (**qdr**), the classification accuracy of our MSGD is 13.5%, while GDCAN only achieves 10.8%. The above experimental results demonstrate the effectiveness of our method for UDA challenge on large-scale dataset.



(a) Accuracy of ablation study

(b) Training stability

(c) Parameter analysis

Figure 3.6: Performance Analysis. (**a**): ablation study on two tasks. (**b**): Training stability. (**c**): Parameter analysis α.

3.3.4 Empirical Analysis

Effect of Intermediate Domain. Based on the comparison result, the core of MSGD is to generate novel instances towards source-like samples. Since no empirical study exists about the working mechanism of MSGD on linking source and target domains, and thus, we attempt to answer such a question by investigating the training process with t-SNE tool. The observations are selected from the tasks Ar→Cl and Pr→Cl. To clearly analyse model behaviors, we first draw the feature visualization of source and target domains at the 10-th epoch and then show the relationship between generative domain and source or target domain at the final stage in Figure 3.4.

From the Figure 3.4 (a) and (d), the learned source features form several clusters with explicit inter-class distance. However, domain shift still has a negative influence on the distribution of target domain, which causes that the trained classifier tends to be invalid for the target object recognition. Thus, our proposed MSGD solves such a challenge from two aspects by using generative instances. The first perspective is to achieve distribution consistence of source and intermediate domains. Our generative strategy creates the corresponding relation between source and novel samples with the same annotation. The class-level alignment not only mitigates their distribution discrepancy but also enhances intra-class compactness as Figure 3.4 (b) and (e). In addition, the diversity of source sample also further facilitates the generalization of model. Another point is to align target and intermediate domains. Since each novel instance is the linear combination of several target samples, it tends to surround target instance providing more contributions to the coding. Compared with source domain, the intermediate domain is more likely to have smaller difference with target domain in Figure 3.4 (c) and (f). To this end, we also achieve domain alignment of source and target domains with the guidance of the intermediate domain.

Ablation Study. We design two variants of our MSGD to explicitly study the effect of each component. One is to remove the CDCT module from our architecture to explicitly show the influence of generative strategy on classification results. The other is to directly replace our MSGD constraint with MMD in the same framework without CDCT module. We carry out experiments on task C→I of Image-CLEF dataset and report experimental performance in Figure 3.6 (a) and the corresponding confusion matrix in Figure 3.5 to delicately observe their difference.

According to Figure 3.6 (a), we notice that the model without CDCT module suffers from significant performance degradation. It demonstrates that the collaborative translation with randomly sampling manner synthesizes several novel instances far from the specific source samples and triggers that MSGD constraint mistakenly measures cross-domain difference. For example, the classifier difficultly distinguishes dog from car by comparing their confusion matrix. The other variant only with MMD is sensitive to environment factors. Specifically, many samples in ImageNet (**I**) come from “bike” category yet usually also include other objects such as “person”. It is hard for model to accurately identify them via MMD constraint. However, the variant MSGD without CDCT module easily overcomes such a problem, which means MSGD loss effectively improves the robustness of model.

Training Stability. Since our method generates novel samples within each mini-batch, a specific source instance is more likely to be represented by the combination of various target samples. It might affect the training stability of the proposed network. Therefore, we further observe the training and test accuracy at each epoch on task **I**→**C** from Image-

CLEF dataset and draw the corresponding curves in Figure 3.6 (b). From the performance, we notice that our method achieves the convergent situation within 20 epochs and there is no considerable change of test accuracy after convergence, which illustrates the whole training process is stable.

Property Analysis. Our model has one trade-off parameter α to control the balance of the discrepancy between source-to-intermediate and target-to-intermediate domains. To analyze its effect, we change α from 0.1 to 0.9 and report how the value of α influences the classification accuracy over two tasks $\mathbf{I} \rightarrow \mathbf{P}$ (Image-CLEF) and task $\mathbf{Ar} \rightarrow \mathbf{Cl}$ (Office-Home) in Figure 3.6 (c). The parameter can adjust the similarity of intermediate domain between source and target domains to affect the trade-off of distribution alignment. Thus, in practice, we select the optimal α and model by evaluating the well-trained model with the sum of source classification error and $\mathcal{A}$-distance.

3.4 Conclusion

Unsupervised Domain Adaptation (UDA) assumes that we have access to well-labeled source data and target instances without annotation, and they share the same label space yet come from different distributions. The main challenge of UDA is to gradually reduce cross-domain discrepancy by learning domain-invariant features. To fight off such a problem, we propose a novel method named Maximum Structural Generation Discrepancy (MSGD) to accurately evaluate source-to-target difference. MSGD involves three important operations. The first task is to construct intermediate domain including synthetic instances from the target domain within each mini-batch with the guidance of source data supervision. It is noteworthy that each generative instance is corresponding to the specific source sample due to structural knowledge. Secondly, we separately adopt class-level and domain-level alignments to eliminate source-to-intermediate and target-to-intermediate discrepancies. The final operation is developing class-driven collaborative translation module to improve the quality of synthetic instance. Extensive experimental results on four challenging visual datasets verify the effectiveness of our MSGD on achieving domain adaptation.

Chapter 4

Dual-Classifier Adversarial Learning for Source-Free Domain Adaptation

4.1 Background

Recent years witness great promising achievements from the exploration of deep neural network (DNN) in the practical scenarios, i.e., image classification, segmentation and detection [110–113]. However, DNN model easily suffers from severe performance degradation when evaluated on test data (target domain) lying in different distribution from the training instances (source domain). Such discrepancy termed as domain shift [96, 114] results from the varying environments or devices [68] and various image styles [115]. To tackle the challenge, unsupervised domain adaptation (UDA) attracts increasing attention and achieves encouraging results by using deep learning architecture.

Conventional UDA assumes the cross-domain data is available during model training, so that it effectively measures and eliminates the domain discrepancy [60, 116, 117]. Based on this assumption, the mainstream solutions of UDA are roughly divided into two paradigms. One branch attempts to transform source and target data into the high-level feature space with the consistent statistical moments to achieve the alignment of their feature distri-

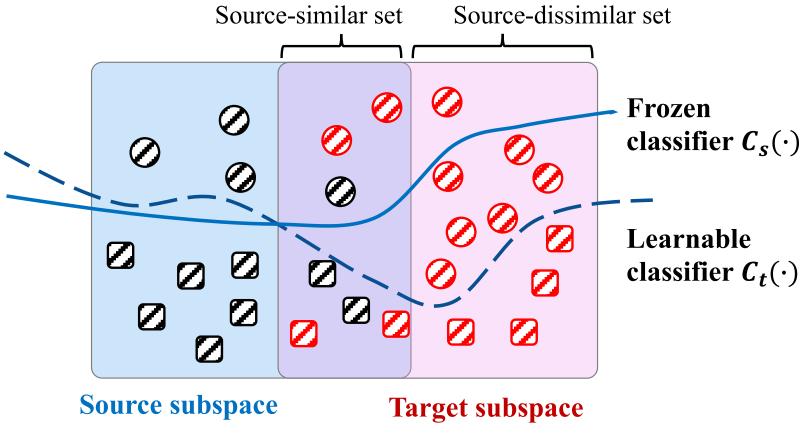


Figure 4.1: Schematic diagram of high-level source (black) and target (red) feature distributions from the trained source model. The target samples can be divided into two subsets: source-similar and source-dissimilar sets. Square and circle represent two different categories. **A^{2}Net** adaptively learns a new classifier (dashed) based on the frozen classifier (solid) trained in source domain.

butions [27, 30, 32, 118, 119]. As for the representative work maximum mean discrepancy (MMD), the learned cross-domain features are enforced to share the identical first-order moment. The other branch devotes more efforts to the deployment of adversarial framework [107, 120, 121]. The core strategy is exploiting feature generator to deceive the domain discriminator so that it fails to recognize which domain the feature comes from. Despite the successes of these methods, it is not hard to observe that they heavily depend on the co-existence of source and target data. However, abundant application scenarios cannot always meet the basic assumption of UDA due to data privacy and memory constraint of small devices. For instance, the training benchmark of ImageNet [11] contains 14 million images occupying hundreds gigabytes storage, which is a huge burden for small-storage equipment. Moreover, many industries such as hospitals are restricted to share their sensitive data with external sites.

The conflict between the practical demand and UDA setting motivates the novel research direction named *Source-free Domain Adaptation* where we are only provided with the well-trained source model instead of well-annotated source data to achieve adaptation to target data. Recently, a few research efforts [42, 44] start exploring this new scenario on cross-domain classification task by assuming that the source classifier contains sufficient knowledge. Thus, they both attempt to directly adjust target features to adapt the source classifier. Among them, SHOT [44], as a simple yet efficient method, freezes the source classifier and integrates pseudo-label supervision and entropy minimization [22] to shorten the distance between target features and source classification boundary. Similarly, MA [42] first considers source classifier as an anchor to guide the generation of new target samples closer to the source domain and then adopts adversarial strategy to achieve domain alignment. In addition, SoFA [122] adopts self-supervised reconstruction to extract more discriminative knowledge from target images themselves to improve the classification ability of model. However, when the data in source domain is imbalanced or insufficient, the above methods with frozen classifier becomes vulnerable due to the lower generalization of source classifier. It is difficult for these approaches to move abundant target features with large variance into the small source classification boundary. For example, as illustrated in Figure 4.1, the source classifier (solid line) trained on the imbalanced data where circle class has only a few data points. Restricted by the frozen classifier, this, unfortunately, leads to a bad classification performance in source-dissimilar set. From another perspective, we post a question: “Can we seek a novel target-specific classifier during model optimization and adapt it to the target features?”.

Along with such a question, we propose a novel Adaptive Adversarial Network (**A^{2}Net**) to address the Source-Free Domain Adaptation. To achieve flexible adjustment for classifier and preserve the original source knowledge, our work firstly introduces a novel target classifier and then exploits dual-classifier design to achieve adversarial domain-level alignment and contrastive category-wise matching (CCM). Concretely, according to the predictions of source and target classifiers, we adaptively divide target samples into two categories: source-similar set and source-dissimilar one. By building such an adversarial relation between dual-classifier and feature generator, **A^{2}Net** gradually eliminates the significant difference across source-similar and source-dissimilar sets and remedies the defect of the frozen source classifier by updating the target classifier. To further learn discriminative features, our work considers the relation of paired samples consisting two any target images as three levels: positive, uncertain and negative pairs, and develops contrastive category-wise matching over all positive pairs to intensify their association. The main contributions of our work are summarized as three folds:

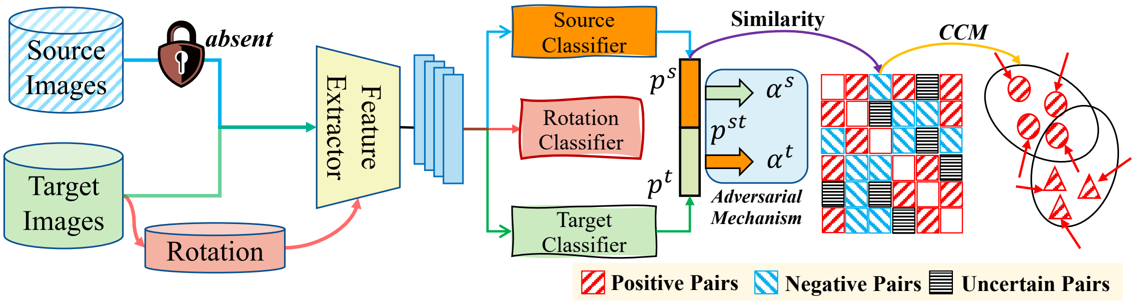


Figure 4.2: Overview of our Adaptive Adversarial Network (**A²Net**) on solving source-free domain adaptation. Given the trained source model including feature extractor $F(\cdot)$ and source classifier $C_s(\cdot)$, we transfer it to identify the target images without source data. To address such a challenge, **A²Net** first adaptively distinguishes source-similar target samples from source-dissimilar ones, and adopts soft-adversarial manner with the introduced target classifier to eliminate their discrepancy. Second, our method explores the contrastive category-wise matching (CCM) to reinforce the relation of positive paired samples. Third, **A²Net** exploits self-supervised rotation to learn more robust and discriminative features.

- First, the proposed **A²Net** integrates a new flexible classifier to be available for optimization and the frozen source classifier to form the dual-classifier architecture which we use to adaptively distinguish source-similar target samples from source-dissimilar ones and achieve alignment across them.

- Second, **A²Net** learns robust and discriminative features in a self-supervised learning manner. Specifically, the contrastive category-wise matching module relies on source knowledge to explore the association of the paired target features and enforce the positive relation to achieve category-wise alignment.

- Finally, we further enhance the model to learn additional semantics through a self-supervised rotation. Experimental results on three benchmarks fully verify the effectiveness of **A²Net** for source-free scenario.

4.2 The Proposed Method

4.2.1 Preliminaries

Given the well-annotated source domain $\mathcal{D}_s = \{(x_i^s, y_i^s)\}$ and unlabeled target instances $\mathcal{D}_t = \{x_i^t\}$, the conventional UDA methods [66, 86, 123, 124] attempt to eliminate domain discrepancy by training a model with the available cross-domain data. However, the practical applications sometimes are restricted to access to original source raw data due to data privacy and/or memory constraint of small devices, which motivates the more challenging Source-Free Domain Adaptation. When adapting to target domain in the novel scenario, we only deploy well-trained source model including feature extractor $F(\cdot)$ and classifier $C_s(\cdot)$ to recognize target samples without any source instances for explicit cross-domain alignment. With the main exploration on how to adapt model to target classification task, our work follows the protocol [44] to train source model by optimizing the $F(\cdot)$ and $C_s(\cdot)$ with the supervisor of source annotation and neglect the specific description on this part.

4.2.2 Adaptive Adversarial Network

From the investigation of Figure 4.1, the considerable domain discrepancy results in the mismatch of feature distribution across source and target domains. Fortunately, there exist some ready-to-recognize target instances similar to source domain distribution for each category. Thus, target high-level features can be divided into two types: source-similar features and source-dissimilar ones. The source classifier $C_s(\cdot)$ confidently identifies source-similar samples. However, it difficultly provides accurate labels for the remaining, especially when trained on insufficient data in source domain. Under such condition, the frozen source classifier in SHOT [44] becomes invalid for the classification of source-dissimilar features, since it is difficult to adapt abundant target features with large variance to $C_s(\cdot)$ with the lower generalization. To avoid the defect, we propose a novel method named Adaptive Adversarial Network (**A^{2}Net**) in Figure 4.2 which alternatively develops a learnable classifier $C_t(\cdot)$ to adapt target feature distribution. The introduced target classifier not only should accurately identify source-similar target feature as C_s but also improves the recognition ability on source-dissimilar ones. Along with the mentioned expectation, the first challenge is to distinguish source-similar features from source-dissimilar ones. However, it is non-trivial to make the decision due to the difficulty of measuring distance between data points and class boundary in high-dimensional feature space.

Soft-Adversarial Inference. Motivated by the voting strategy, we compare the output of classifiers to adaptively determine the type of features. Specifically, each target sample through two classifiers in Figure 4.2 achieves its probability distribution of category before Softmax operation $p_i^s = C_s(F(x_i^t)) \in \mathbb{R}^K$ and $p_i^t = C_t(F(x_i^t)) \in \mathbb{R}^K$, where K is the number of class. Subsequently, we activate the concatenation of p_i^s and p_i^t with Softmax function $\sigma(\cdot)$ to access $p_{(i)}^{st} = \sigma([p_i^s \; p_i^t]^\top) \in \mathbb{R}^{2K}$, and consider $\alpha_i^s = \sum_{k=1}^{K} p_{(i)k}^{st}$ and $\alpha_i^t = \sum_{k=K+1}^{2K} p_{(i)k}^{st}$ as voting scores. When α_i^s is larger than α_i^t, the corresponding feature belongs to source-similar set, otherwise, it is divided into the other subset. The definition gives us a manner to optimize target classifier and feature extractor with:

$$\begin{aligned} \min_{F,C_t} \quad & -\textstyle\sum_{i=1}^{n_t} \mathbb{I}(\alpha_i^s > \alpha_i^t)\sigma(p_i^s)\log(\sigma(p_i^s)) \\ & -\textstyle\sum_{i=1}^{n_t} \mathbb{I}(\alpha_i^s \leq \alpha_i^t)\sigma(p_i^t)\log(\sigma(p_i^t)), \end{aligned} \tag{4.1}$$

where $\mathbb{I}(\cdot)$ is the indicator function. However, such a constraint easily gives rise to the necessary concern "What will happen if $C_t(\cdot)$ generates wrong prediction when $\alpha_i^s \leq \alpha_i^t$?" Under this situation, the prediction tends to be far away from the ground-truth. Thus, the trade-off between accepting novel target knowledge and preserving well-learned source knowledge becomes important, and we further rewrite Eq. (4.1) as:

$$\mathcal{L}_c = -\sum_{i=1}^{n_t} \Big(\alpha_i^s \sigma(p_i^s) \log\big(\sigma(p_i^s)\big) + \alpha_i^t \sigma(p_i^t) \log\big(\sigma(p_i^t)\big) \Big),$$

where α_i^s and α_i^t are frozen during optimization.

From another perspective, we also consider the source-similar and source-dissimilar high-level features distributing in two independent domains. The alignment of them further reduces their discrepancy to learn more discriminative features. In addition, the introduced target classifier $C_t(\cdot)$ finally has the equivalent classification ability for source-similar ones. According to the dual-classifier design, we propose a **Soft-Adversarial** mechanism to address the above demands with the formal objective function as:

$$\begin{aligned} \min_{C_t} \mathcal{L}_{c'} &= -\sum_{i=1}^{n_t} (\alpha_i^s \log(\sum_{k=1}^{K} p_{(i)k}^{st}) + \alpha_i^t \log(\sum_{k=K+1}^{2K} p_{(i)k}^{st})), \\ \min_{F} \mathcal{L}_{c''} &= -\sum_{i=1}^{n_t} (\alpha_i^t \log(\sum_{k=1}^{K} p_{(i)k}^{st}) + \alpha_i^s \log(\sum_{k=K+1}^{2K} p_{(i)k}^{st})). \end{aligned}$$

To explicitly understand the Soft-Adversarial loss, we firstly illustrate that $\alpha_i^{s/t}$ denotes the probability of the sample x_i belonging to source-similar or source-dissimilar subsets and $\alpha_i^s + \alpha_i^t = 1$. For the extreme condition such as $\alpha_i^s \approx 1$, the optimization of $\ell_{c'}$ further reduces the discriminability of target classifier $C_t(\cdot)$ for x_i. However, feature generator engages in the inverse operation mapping x_i into high-level representation similar to source-dissimilar part by minimizing $\ell_{c''}$. Beneficial from the adversarial manner between feature generator and classifiers, we further align source-closer and source-dissimilar sets and eliminate the difference of classifiers.

Contrastive Category-wise Matching. The core motivation of adaptive adversarial inference is to discover source-similar target samples and achieve domain-level alignment across source-similar and source-dissimilar sets. However, domain-invariant features learned with adversarial learning fail to represent the category-level matching. In addition, without annotation over target domain, it becomes difficult to identify the category relationship among samples. The intuitive solution to the challenge is to directly provide each sample with pseudo-label, which easily results in the negative influence on model training, especially during the initialization stage, due to the uncertainty of pseudo-label. Inspired by the contrastive learning [125], we design a novel discriminative dual classifier exploring the association of paired samples to achieve the class-wise alignment in unsupervised manner.

Concretely, each visual instance within a batch is transformed into the label space via the source classifier $\mathcal{I}_i = \sigma(p_i^s) \in \mathbb{R}^K$ which we use to capture the similarity of any paired samples through $s_{ij} = \mathcal{I}_i^\top \mathcal{I}_j$ in Figure 4.2. The larger s_{ij} denotes that the i-th and j-th samples belong to the same category with higher probability, vice versa. However, we fail to confidently judge the relationship of several pairs when s_{ij} lies in the middle interval. Thus, all pairs of each mini-batch are divided into three subsets: positive, uncertain and negative sets by comparing s_{ij} with the upper bound $\mu(t)$ and lower bound $\ell(t)$ defined as:

$$\begin{cases} \mu(t) = \mu_0 - \lambda_\mu \cdot t \\ \ell(t) = \ell_0 + \lambda_{\ell \cdot t} \\ 0 \leq \ell(t) \leq \mu(t) \leq 1 \end{cases} \qquad \gamma_{ij} = \begin{cases} 1, & s_{ij} > \mu(t) \\ -1, & s_{ij} < \ell(t) \\ 0, & otherwise \end{cases}$$

where $\mu(t)$ as well as $\ell(t)$ are the linear functions of epoch t starting from zero, and μ_0 and ℓ_0 are the initial upper and lower bounds, respectively, and λ_μ and λ_ℓ separately control the decreasing and increasing rate of μ_0 and ℓ_0. The pairs are definitely classified into positive ($\gamma_{ij} = 1$) and negative ($\gamma_{ij} = -1$) subsets when $s_{ij} > \mu(t)$ and $s_{ij} < \ell(t)$, respectively. For other cases, we temporarily neglect the ambiguous associations with $\ell(t) < s_{ij} < \mu(t)$ by $\gamma_{ij} = 0$. As the piecemeal change of $\mu(t)$ and $\ell(t)$, our method adaptively makes the judgement for more pairs.

To achieve class-wise alignment, the relation of positive pairs must be further intensified to learn more similar feature representation for themselves. Similar with [125], we develop the contrastive loss for each positive pair of example (i, j) formulated as:

$$\xi(i,j) = -\log \frac{\exp(s_{ij})}{\sum_{v=1}^{b} \mathbb{I}(v \neq i)|\gamma_{iv}| \exp(s_{iv})}, \tag{4.2}$$

where b is the size of each batch and $|\gamma_{iv}|$ means the absolute value of γ_{iv}. According to the monotonic property[1], we obtain the optimization of Eq. (4.2) approximates the minimum value of function with $s_{ij} \to 1$. That illustrates $\sigma(p_i^s)$ and $\sigma(p_j^s)$ follow the more similar probability distribution. And the property is transmitted into the output of feature generator due to the frozen source classifier so that samples from the identical class distribute closer to each other in the high-level feature space. Thus, we adopt Eq. 4.2 on all positive pairs and reformulate the final contrastive objective:

$$\min_F \mathcal{L}_p = \mathbb{I}[\mu(\lambda) > \ell(\lambda)] \sum_{i=1}^{b} \sum_{j=1, j\neq i}^{b} \mathbb{I}(\gamma_{ij} = 1)\xi(i,j). \tag{4.3}$$

Note that Eq. (4.3) makes no sense under $\mu(\lambda) \leq \ell(\lambda)$ since we fail to find new positive pairs to optimize it.

Self-Supervised Rotation. So far, we mainly consider how to transfer knowledge into target domain only with the guidance of well-trained source model. However, the pure classification model heavily relies on the given source data, which is often lying imbalanced distribution, further limiting the generalization ability of target classifier. To solve this, we

[1] When the variable $x \in [0,1]$, the objective function $f(x) = -\log \frac{\exp(x)}{\exp(x)+a}$ achieves the minimum value when $x = 1$, since $f^{'}(x) < 0$ meaning $f(x)$ is monotonically decreasing when $x \in [0,1]$.

explore self-supervised rotation manner over target domain to augment the sample space, which enhances the learning of feature extraction and target classifier. In other words, the model is able to easily see more variances per high-confident predicted target sample. Following [126], we set four rotation degrees $\theta \in \{0^o, 90^o, 180^o, 270^o\}$ with corresponding 4-class rotation labels y_r. Within one batch, we randomly select rotation label y_r and then have access to the new processed image $\hat{x}_i^t$ by rotating the original image x_i^t with $90^o y_r$. In addition, we also introduce the rotation classifier $C_r(\cdot)$ in Figure 4.2 taking $F(\hat{x}_i^t)$ as input and predicting the rotation label. Finally, the cross-entropy loss is adopted to measure the difference between prediction and ground-truth rotation as follows:

$$\min_{F,C_r} \mathcal{L}_r = -\sum_{i=1}^{b} y_i^r \log(F(\hat{x}_i^t)). \tag{4.4}$$

By identifying the rotation degree, the model effectively captures the important visual signals from original images for object classification.

4.2.3 Overall Objective and Optimization

The above description has specifically illustrated how our method works for source-free domain adaptation. It is simple to notice that the training of model mainly involves the update of three modules (i.e., feature generator $F(\cdot)$, rotation classifier $C_r(\cdot)$, and target classifier $C_t(\cdot)$) with the following overall objective:

$$\min_{F,C_r} \mathcal{L}_c + \mathcal{L}_{c''} + \mathcal{L}_p + \eta\mathcal{L}_r, \tag{4.5}$$

$$\min_{C_t} \mathcal{L}_c + \mathcal{L}_{c'}, \tag{4.6}$$

where η is the trade-off parameter. To achieve the adaptive adversarial operation, we adopt iterative manner to alternately optimize three modules. **First**, the source and target classifiers take the features from generator as input to access the class prediction which we use to update the feature generator and rotation classifier via Eq. (4.5) with frozen the target multi-class classifier $C_t(\cdot)$. **Second**, we only optimize the target classifier when fixing $F(\cdot)$ and $C_r(\cdot)$ with Eq. (4.6). **Third**, the adversarial training repeats the above two steps until we reach the convergence or maximum epochs.

4.3 Experiments

4.3.1 Experimental Details

Datasets: In experiments, we evaluated our proposed method and other baselines on three cross-domain benchmarks, i.e., Office-31, Office-Home and VisDA. The descriptions of Office-31 and Office-Home are in Chapter 2 and the introduce of VisDA is in Chapter 3.

Implementation: According to [44], for the source model, we separately consider Resnet-50 and Resnet-101 as backbones to extract high-level features from two object datasets and VisDA, and replace the original last FC layer with a new bottleneck layer followed by Batch Normalization (BN). The source classifier C_s consists of one FC layer and a weight normalization layer. During adaptation, we introduce the target classifier C_t with the same architecture as C_s and the rotation classifier C_r including two FC layers. In addition, we initialize C_t with the parameters of C_s by appending Gaussian noises from $N(0, I)$. As for the optimizer, we adopt SDG with momentum 0.9 and weight decay $1e^{-3}$. The initial learning rates on Office-31/Office-Home for the pre-trained backbone and new added components are $1e^{-3}$ and $1e^{-2}$ respectively, however, they are set as $1e^{-4}$ and $1e^{-2}$ for VisDA. Moreover, we set the identical upper and lower bounds for all experiments as $\mu_0 = 0.95$, $\ell_0 = 0.45$, $\lambda_\mu = 9.9e^{-3}$ and $\lambda_\ell = 9.9e^{-4}$.

Baselines: The comparisons include two categories of domain adaptation algorithms. One is vanilla domain adaptation, which requires source and target data at the same time to solve the domain shift, such as Resnet [76], DANN [36], SAFN [83], CDAN [19], SRDC [20], BNM [127] and MCC [128]. Additionally, we also compare the recent state-of-the-art source-free domain adaptation models, i.e., SFDA [129], SHOT [44], SDDA [43] and SoFA [122]. Note that since MA [42] needs to generate additional target samples on solving source-free task, we make no comparison with it.

Table 4.1: Comparisons of Object Classification Accuracy (%) of Source-free Domain Adaptation on Office-31. The best accuracy for source-free tasks is highlighted with **bold** type, while we use underline to emphasize the highest result for source-need task.

	Method	A→D	A→W	D→A	D→W	W→A	W→D	Avg
Source-Needed	ResNet [76]	68.9	68.4	62.5	96.7	60.7	99.3	76.1
	DANN [36]	79.7	82.0	68.2	96.9	67.4	99.1	82.2
	SAFN [83]	90.7	90.1	73.0	98.6	70.2	99.8	87.1
	CDAN [19]	92.9	94.1	71.0	98.6	69.3	100.0	87.7
	BNM [127]	90.3	91.5	70.9	98.5	71.6	100.0	87.1
	MCC [128]	95.6	95.4	72.6	98.6	73.9	100.0	89.4
	SRDC [20]	95.8	95.7	76.7	99.2	77.1	100.0	90.8
Source-Free	SFDA [129]	92.2	91.1	71.0	98.2	71.2	99.5	87.2
	SDDA [43]	85.3	82.5	66.4	99.0	67.7	99.8	83.5
	SoFA [122]	73.9	71.7	53.7	96.7	54.6	98.2	74.8
	SHOT [44]	94.0	90.1	74.7	98.4	74.3	99.9	88.6
	Ours	**94.5**	**94.0**	**76.7**	**99.2**	**76.1**	**100.0**	**90.1**

4.3.2 Comparison Results

Tables 4.1-4.3 report the results of object classification on Office-31, Office-Home and VisDA, respectively. From the investigation of them, our proposed Adaptive Adversarial Network (**A^{2}Net**) achieves the highest average accuracy across three benchmarks when compared with others for source-free domain adaptation, which illustrates the design of **A^{2}Net** effectively transfers knowledge only from source model to assist the target data recognition. In addition, we can easily achieve three important conclusions by making delicate comparisons over these competitors.

First of all, **A^{2}Net** provides well-trained source model with more powerful adaptation ability when evaluated on unsupervised target domain, especially for small scale source domain. For example, with **D** and **W** as source domains on Office-31, our approach outperforms the second highest accuracy from SHOT by 2.0% and 1.8% when adapting to the target domain **A**. As we all know, there exists serious imbalanced data scale challenge

Table 4.2: Comparisons of Object Classification Accuracy (%) of Source-free Domain Adaptation on Office-Home. The best accuracy for source-free tasks is highlighted with **bold** type, while we use underline to emphasize the highest result for source-need task.

	Source	Art (Ar)			Clipart (Cl)			Product (Pr)			Real-World (Rw)			Avg
	Target	Cl	Pr	Rw	Ar	Pr	Rw	Ar	Cl	Rw	Ar	Cl	Pr	
Source-Needed	Resnet [76]	46.3	67.5	75.9	59.1	59.9	62.7	58.2	41.8	74.9	67.4	48.2	74.2	61.3
	DANN [36]	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
	SAFN [130]	52.0	71.7	76.3	64.2	69.9	71.9	63.7	51.4	77.1	70.9	57.1	81.5	67.3
	CDAN [19]	50.7	70.6	76.0	57.6	70.0	70.0	57.4	50.9	77.3	70.9	56.7	81.6	65.8
	BNM [127]	52.3	73.9	80.0	63.3	72.9	74.9	61.7	49.5	79.7	70.5	53.6	82.2	67.9
	SRDC [20]	52.3	76.3	81.0	69.5	76.2	78.0	68.7	53.8	81.7	76.3	57.1	85.0	71.3
Source-Free	SFDA [129]	48.4	73.4	76.9	64.3	69.8	71.7	62.7	45.3	76.6	69.8	50.5	79	65.7
	SoFA [122]	-	74.1	77.6	-	71.8	75.1	-	-	-	-	-	-	-
	SHOT [44]	57.1	78.1	81.5	**68.0**	78.2	78.1	67.4	54.9	82.2	73.3	58.8	84.3	71.8
	Ours	**58.4**	**79.0**	**82.4**	67.5	**79.3**	**78.9**	**68.0**	**56.2**	**82.9**	**74.1**	**60.5**	**85.0**	**72.8**

across source and target domain, i.e., **D** (498) vs **A** (2,817) and **W** (795) vs **A**. The classifier firstly trained on small-scale source domain has so insufficient generalization ability that it ineffectively is applied to large-scale target domain. Thus, it is difficult for SHOT with frozen classifier to accurately move abundant target features into the source classification boundary. However, our **A^{2}Net** adopts flexible target classifier with adversarial training to adapt it to target features. This is the main reason for our success on these two tasks. And **A^{2}Net** beats several UDA based methods such as SAFN and BNM by a large margin, which means even if we fail to access the source data, our method still exploits the finite source knowledge to achieve better adaptation.

Second, our proposed method also effectively overcomes the negative influence of significant domain discrepancy. To the best of our knowledge, there exists significant domain shift between **Ar** and **Cl** because of the considerable difference of image styles. However, **A^{2}Net** surpasses SFDA by 10% for this adaptation task since our proposed method adaptively distinguishes the source-similar target samples from source-dissimilar ones and explores adversarial manner to gradually eliminate domain discrepancy. Moreover, with the

Table 4.3: Comparisons of Object Classification Accuracy (%) of Source-free Domain Adaptation on VisDA. The best accuracy for source-free tasks is highlighted with **bold** type, while we use underline to emphasize the highest result for source-need task.

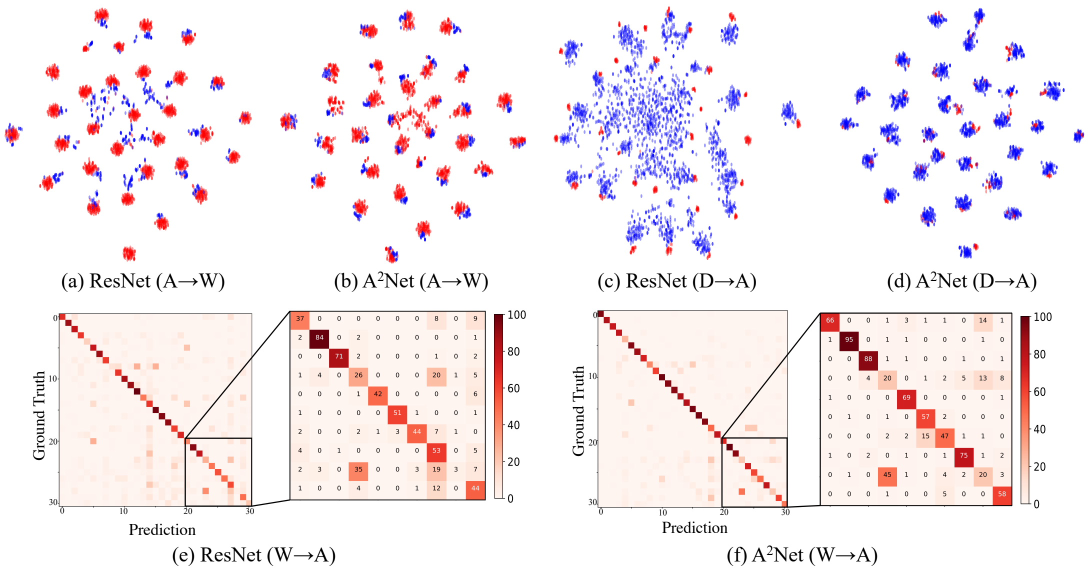
	Methods	plane	bcycl	bus	car	horse	knife	mcycl	person	plant	sktbrd	train	truck	Per-Class
Source-Needed	Resnet [76]	55.1	53.3	61.9	59.1	80.6	17.9	79.7	31.2	81.0	26.5	73.5	8.5	52.4
	DANN [36]	81.9	77.7	82.8	44.3	81.2	29.5	65.1	28.6	51.9	54.6	82.8	7.8	57.4
	CDAN [19]	85.2	66.9	83.0	50.8	84.2	74.9	88.1	74.5	83.4	76.0	81.9	28.0	73.9
	SAFN [83]	93.6	61.3	84.1	70.6	94.1	79.0	91.8	79.6	89.9	55.6	89.0	24.4	76.1
	MCC [128]	88.7	80.3	80.5	71.5	90.1	93.2	85.0	71.6	89.4	73.8	85.0	36.9	78.8
Source-Free	SFDA [129]	86.9	81.7	84.6	63.9	93.1	91.4	**86.6**	71.9	84.5	58.2	74.5	42.7	76.7
	SoFA [122]	-	-	-	-	-	-	-	-	-	-	-	-	64.6
	SHOT [44]	**94.3**	**88.5**	80.1	57.3	93.1	94.9	80.7	80.3	91.5	**89.1**	86.3	**58.2**	82.9
	Ours	94.0	87.8	**85.6**	**66.8**	**93.7**	**95.1**	85.8	**81.2**	**91.6**	88.2	**86.5**	56.0	**84.3**

increasing of object category, we notice that all methods suffer from the performance degradation on Office-Home when compared with their results of Office-31. But the contrastive category-wise matching depends on the constraint over positive paired samples to learn so explicit classification boundary that $\mathbf{A^2}$**Net** still achieves the best performance.

Third, the experimental results in Table 4.3 fully demonstrate that our designed algorithm makes sense to solve source-free domain adaptation with the large-scale benchmark. Specifically, $\mathbf{A^2}$**Net** obtains higher classification accuracy than other state-of-the-art methods in most adaptation tasks on VisDA and makes more accurate identify on several confusing objects such as bus and car.

4.3.3 Empirical Analysis

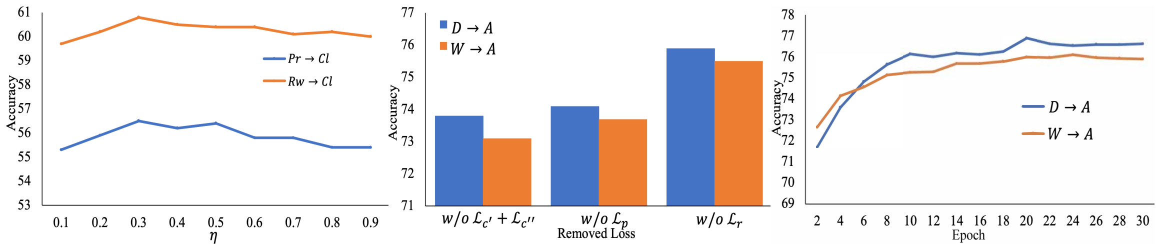
Feature Visualization & Confusion Matrix: According to the experimental results in Table 4.1 and the working mechanism of model, when compared with other state-of-the-art baselines, it is simple to notice that $\mathbf{A^2}$**Net** is non-sensitive to the mismatch of cross-domain data scale, where source domain contains much fewer instances than target domain. To further explore how our work achieves it, we provide the visualization of embedding feature and confusion matrix in Figure 4.3 with large- or small- scale source domain. Concretely,



(a) ResNet (A→W)
(b) A^{2}Net (A→W)
(c) ResNet (D→A)
(d) A^{2}Net (D→A)
(e) ResNet (W→A)
(f) A^{2}Net (W→A)

Figure 4.3: Resulst of Feature Visualization and Confusion Matrix. (a)-(d) show high-level source (red) and target (blue) features generated by source-only model (Resnet-50) and our **A^{2}Net**. Note that we only exploit source data to draw the t-SNE without any use of it during adaptation stage. (e) and (f) are the confusion matrices, comparing the ground-truth and the category prediction from ResNet and our model, respectively.

the well-trained target model of **A^{2}Net** and source-only ResNet are frozen to extract the high-level features before the classifier from the unseen source domain and unlabeled target one. And we carry out the experiments on Office-31 since it exists the imbalanced data scale challenge, i.e., **A** (2,817) vs **D** (498), and **A** (2,817) vs **W** (795). The comparison between Fig. 4.3 (a) and (b) illustrates the model trained on large-scale source domain has more powerful generalization ability than that with smaller-scale one. With **A** as source domain, **A^{2}Net** easily distinguishes source-similar target features from source-dissimilar ones and gradually aligns these two parts by using soft-adversarial mechanism. Thus, after adaptation, target features of each category (produced by **A^{2}Net**) almost distribute the boundary of source domain. Under this condition, our target classifier being similar to the original source one exactly identifies them. However, with source model trained on **D**, even if the model has finished the adaptation, there are abundant target features far away from the corresponding source class so that the original source classifier difficultly make an accurate decision on them. The flexible target classifier of our method, thus, fully shows the importance of its optimization, which facilitates model to adapt itself to target features



(a) Parameter Anslysis (b) Ablation Study (c) Training Stability

Figure 4.4: (a) Parameter Analysis records the object recognition accuracy with the varying η. (b) Ablation Study shows the influence of removing each constraint on the performance of our model. (c) Training Stability reports the object recognition ability of target classifier as the increasing number of epoch.

instead of only adjusting feature learning as SHOT [44]. Beneficial from the dual-classifier design, our work achieves the highest classification accuracy on task **D→A** in Table 4.1. And, the confusion matrix derived from task **W→A** demonstrates our method learns more compact category subspace by intensifying the association of positive pairs with contrastive loss to achieve category-wise matching across source-similar and source-dissimilar sets.

Ablation Study, Parameter Analysis & Training Stability: Our **A^{2}Net** mainly consists of three modules: soft-adversarial inference, contrastive category-wise matching and self-supervised rotation which support the model adaptation from various perspectives. Therefore, we attempt to separately remove each module from them to investigate the change of classification accuracy on two tasks **D→A** and **W→A**. According to the experimental results in Fig. 4.4 (b), we achieve the conclusion that the soft-adversarial mechanism has an important and positive influence on improving the generalization of model. Without the adversarial operation, it becomes tough to effectively promote adaptation of the target classifier so that the model heavily relies on the performance of the frozen source classifier. Similarly, removing the contrastive category matching also results in the performance degradation since this module mainly exploits the existed knowledge of source model to explore the relation of any two target samples and controls the compactness of each target class subspace by using contrastive loss over all positive pairs. In terms of the rotation design, it actually makes a small contribution to the improvement of performance by learning additional semantics from target images in self-supervised manner. However, we still promote the adaptation ability of model via the adjustment of parameter η balancing the rotation constraint and others. For instance, Fig. 4.4 (a) reports the relation between the varying η and target classification accuracy. These two tasks of Office-Home both achieve the highest performance with $\eta = 0.3$. Finally, considering the adversarial game between feature generator and dual-classifier , we show the change of object recognition accuracy as the increasing of epoch in Fig. 4.4 (c). With the adversarial training manner, the target classifier gradually improves its classification ability in a stable rhythm.

4.4 Conclusions

Unsupervised Domain Adaptation (UDA) assumes the well-annotated source domain and unlabeled target images are both available for the model training. However, many practical applications only access the well-trained source model instead of source data during adaptation stage, which is defined as source-free domain adaptation. To overcome the novel scenario, this paper proposes Adaptive Adversarial Network (**A^{2}Net**) including three operations. First, **A^{2}Net** develops a soft-adversarial mechanism to learn a flexible target classifier promoting the recognition of samples which the frozen source classifier difficultly identifies. Second, it explores the contrastive loss over all positive paired target samples to intensify the compactness of each category subspace. Finally, the self-supervised rotation is adopted to learn additional semantics from target images to learn more discriminative features. Moreover, experiments of three popular benchmarks illustrate our method effectively achieves domain adaptation without source data.

Chapter 5

Representation Generation for Imbalanced Domain Generalization

5.1 Background

Deep learning recently achieves great success in various learning tasks, e.g., object recognition [131, 132], semantic segmentation [133] and image generation [134]. Such achievements typically benefit from extensive well-annotated instances in the training stage. However, these favorable conditions hardly occur in the reality, especially for the situation where samples are collected from different environments or devices. These differences are likely to result in the data distribution shift. Hence, the well-trained model encounters with significant performance degradation when assessing it on test set [135].

This challenge motivates many research efforts [20, 22] on unsupervised domain adaptation (UDA), where most of them attempt to mitigate domain shift by learning domain-invariant features given label-sufficient source domain and unlabeled target instances [136, 137]. The pre-requisite of domain alignment is accessing target samples during the training stage to measure and minimize cross-domain discrepancy. However, the collection of abundant target instances becomes expensive and laborious. For example, police generally only have the sketch image of victims at hand to identify victims from RGB-based photos captured by the widely-used surveillance system. Thus, learning a high-generalization

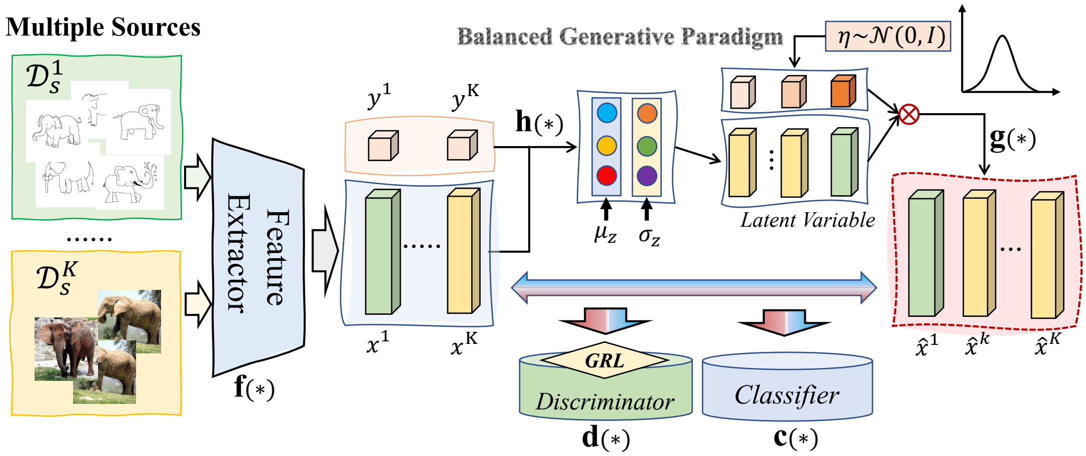


Figure 5.1: Overview of the proposed architecture, where the feature extractor $\mathbf{f}(\cdot)$ maps multiple source images into hidden representations $\mathbf{x}$ which are used to estimate joint distribution and also fed into multi-class classifier $\mathbf{c}(\cdot)$ and multi-domain discriminator $\mathbf{d}(\cdot)$. Moreover, the balanced generative paradigm (BGP) aims to augment sufficient novel samples to improve model generalization via the balance of data scale. Concretely, BGP first utilizes the network $\mathbf{h}(\cdot)$ to estimate the statistics (μ_z and σ_z) of latent variables deriving cross-domain images from the same category and the other component $\mathbf{g}(\cdot)$ relies on the learned variables to synthesize novel samples. These generative instances will facilitate discriminator and classifier to be more robust and discriminative.

classification model to identify unseen target images becomes very essential.

Motivated by this demand, domain generalization (DG) attracts extensive attentions in recent years [138, 139], which aims to capture domain-invariant knowledge from the collaboration of multiple source domains and identify unseen target images during evaluation stage [49, 115, 140]. Due to the success of adversarial learning, [50] explored this strategy to extract transferable representations over source domains under DG scenario. In addition, to roughly estimate properties of the unseen target domains, [47, 48, 141] integrated images from various domains to synthesize novel visual signals to further promote the generalization of feature extractor. Inspired by jigsaw puzzle, another branch for DG seeks the intrinsic semantic knowledge of object by mining the association among image patches with self-supervised loss [53, 54]. Meta-learning based approaches explore episodic training manner to overcome DG issue [142]. Similarly, [139] gradually activated neurons related to domain-invariant semantics by adjusting their gradient and [49] utilized Fourier transformation to filter out the essential knowledge from visual signals for the downstream tasks as data augmentation and classification.

Although the existing DG works achieve remarkable performance, they mainly neglect the negative effect of imbalanced data scale across source domains and category, especially for methods based on adversarial training. These solutions generally seek a discriminator to identify which domains each sample comes from and confuse feature extractor to generate domain-invariant features. Under this condition, the discriminator has naturally become a multi-domain classifier. When there exist imbalanced data scale across source domains, the normal classification issue also becomes a akin long-tail distribution recognition problem [143, 144]. Training discriminator under this situation will reduce its discriminative ability and robustness, which indirectly yields negative influence on learning transferable representations. On the other hand, the imbalanced category distribution per source domain also hinders the learning of robust domain-invariant features. The supervised learning on source domains with long-tail category distribution is likely to facilitate classifier to yield predictive bias. In other words, the classifier easily recognizes samples of majority classes while difficultly determines the categories of instances from minority ones. The less robust classification model fails to be well generalized into the usage of target domain. Moreover, imbalanced data scale issue is widespread in the practical applications. For instance, in the popular benchmark PACS [51], the sample number **S**ketch domain is much larger than that of others.

In this paper, we formulate the above scenario as a practical and challenging problem *Imbalanced Domain Generalization* (IDG). To solve IDG, the straightforward approach is to compensate enough novel samples into minorities and utilize them to learn robust and discriminative classifier and discriminator. Along with this line, we propose a simple yet effective novel method “Generative Inference Network (GINet)”. As Figure 5.1 shows, our GINet observes the available cross-domain samples from the same category to infer their common latent variable deriving them by removing domain-specific semantics and explores the deduced variables to generate novel reliable and meaningful instances for downstream tasks with optimal transport constraint. Concretely, we first adopt the mature network architecture, e.g., ResNet [76] as backbone followed by classifier and discriminator to extract domain-invariant features. Second, given multiple source features, our balance generative module of GINet deduces the their direct cause to capture the association of different domains and uniformly record attributions of each domain. Finally, the optimal transport

(OT) mechanism is explored to minimize the distribution divergence across novel derived samples and original ones to guarantee high-quality generation. In a nutshell, our principal contributions are summarized as:

- First, our work mainly focuses on the negative effect of imbalanced data scale across domains and categories on learning high-generalization classification model and formulates it as IDG scenario. This unfavorable training condition heavily reduces the robustness of classifier and discriminator and makes the whole system difficult to capture transferable knowledge.

- Second, we develop a simple yet effective novel method "generative inference network (GINet)" to overcome challenges of IDG. Our GINet deduces the latent variable deriving cross-domain images from the identical category and relies on them to generate sufficient novel samples for minorities. Moreover, we explore optimal transport alignment mechanism to achieve high-quality generation.

- Finally, we design the corresponding imbalanced experimental setting over three widely-used benchmarks (PACS, VLCS and Office-Home) for empirical analysis. The extensive experiments and ablation studies comprehensively show the advantage of our GINet over other DG methods on promoting model generalization under IDG scenario.

5.2 The Proposed Method

5.2.1 Motivation and Preliminaries

For domain generalization, we have access to K source domains $\{\mathcal{D}_s^k\}_{k=1}^K$, where the k-th domain $\mathcal{D}_s^k$ involves N_k well-labeled samples $\{(\mathbf{X}_i^k, \mathbf{y}_i^k)\}_{i=1}^{N_k}$, and $\mathbf{y}_i^k \in \{1, 2, \cdots, C\}$ represents the corresponding label. During evaluation stage, the fully-trained model from source domains attempts to recognize unseen target images $\mathcal{D}_t = \{\mathbf{X}_i^t\}_{i=1}^{N_t}$. However, the considerable cross-domain distribution divergence makes the well-trained source model invalid to target domain. The feasible solution [50] is to discover domain-invariant knowledge via the collaborative training over multiple source domains using either adversarial domain training or cross-domain discrepancy minimization.

When there exist domain-level imbalance across source domains, the domain classification also becomes an akin long-tail distribution recognition problem [143, 144]. For adversarial learning, the discriminator as a multi-domain classifier under this situation cannot achieve good discriminative ability and robustness, which indirectly yields negative influence on learning domain-invariant representations. On the other hand, the category-level imbalanced distribution with each source domain also hinders the learning of robust transferable features. In the unfavorable training condition, the classifier easily discovers semantics of samples from majorities and accurately identifies them, while tends to be weak on minorities. For instances, the popular benchmark PACS consists of four various domains showing somewhat imbalanced data scale issue between **S**ketch and **P**hoto domains. When evaluating one recent DG method (DMG) [145] with AlexNet as backbone on PACS dataset, it always achieves better performance on **S**ketch than **P**hoto (92.95% v.s. 89.03%) with **A**rt as target domain. This case indicates that the imbalanced data scale indeed reduces the discriminative ability and fails to capture robust transferable representations from visual images.

In this paper, We formally define a practical and challenging "imbalanced domain generalization (IDG)" problem, which assumes that there exists significant difference of data scale across various source domains and categories. From domain level, the sample number of the smallest source domain tends to be many times smaller than that of the largest one. Without loss of generality, N_k gradually becomes pretty large as the increasing of k. In addition, each source domain also shows imbalanced category distribution. Specifically, for the k-th source domain with $N_k = \sum_{j=1}^{C} n_j^k$ images, the number of samples per category n_j^k shows an obvious increasing trend with larger subscript j. Under IDG setting, the model easily learns more discriminative knowledge from majorities, which has negative influence on learning transferable semantics. To solve challenges of IDG, we propose a novel learning algorithm "generative inference network (GINet)" to ameliorate model generalization via the balance of data scale.

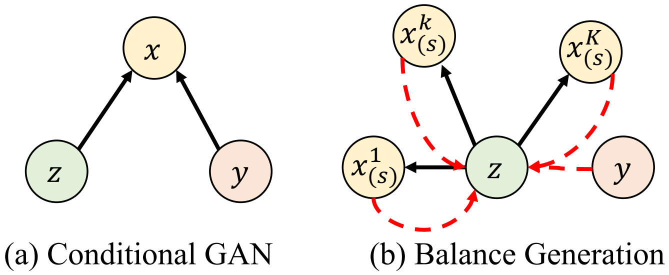


Figure 5.2: Comparison of Generative Theoretical Paradigm.

5.2.2 Balanced Generative Paradigm

The major obstacle for learning high-generalized model under IDG setting becomes the divergence of data scale across source domains and categories. The intuitive solution is to synthesize sufficient reliable samples to compensate those minority domains or classes. One potential strategy is that conditional GAN (cGAN) [37] synthesizes class-specific sample $\mathbf{x}$ conditioned on the combination of random noise ($\mathbf{z} \in \mathbb{R}^d$) and one-hot label ($\mathbf{y} \in \mathbb{R}^C$) in the latent encoding space as Fig. 5.2 (a) shows. However, the smaller source domains or categories with limited training samples fail to provide sufficient knowledge for the generative process during the training stage. Hence, the generator tends to produce more instances similar with majorities to further aggravate the imbalanced situation. More importantly, the random noise is unable to reflect hidden relationship of cross-domain images from the identical category, which negatively affects the generation of the orientation-related features.

To effectively handle the uncertainty from random noise and the partiality of generator, we present a novel Balance Generative Paradigm (BGP) to discover the direct causal deriving cross-domain images with the consistent annotation by eliminating domain-specific semantics and then utilize these latent variables to augment novel samples with more diversities. As Fig. 5.2 (b) shows, our BGP first assumes that cross-domain instances belonging to the same class and their corresponding annotation are most likely derived from the same latent variable $\mathbf{z}$. Therefore, the joint probability distribution over latent variable and the observable samples is formulated as:

$$p(\mathbf{x}^1, \cdots, \mathbf{x}^K, \mathbf{y}, \mathbf{z}) = \prod_{k=1}^{K} p(\mathbf{x}^k|\mathbf{z})p(\mathbf{y}|\mathbf{z})p(\mathbf{z}). \tag{5.1}$$

From the inverse perspective, we expect to deduce the shared direct cause given the available cross-domain samples. Thus, the paradigm is also modeled from the opposite direction where the hidden state $\mathbf{z}$ is inferred through the joint generation over observable source data and annotation, i.e., $p(\mathbf{z}|\mathbf{x}^1, \cdots, \mathbf{x}^K, \mathbf{y})$. According to the intact paradigm, the balanced generation in the practical training mainly involves two phases. First, for each category, we randomly sample points from various source domains with the same quantity to estimate their joint distribution, and derive the latent variable $\mathbf{z}$. Beneficial from such a manner, the latent variable averagely reflects attribution of each source domain and comprehensively captures their hidden correlation facilitating the following generation of domain-invariant features. Second, to effectively overcome imbalanced issue, for each specific domain, the sampling number of latent variable $\mathbf{z}$ from $p(\mathbf{z}|\mathbf{x}^1, \cdots, \mathbf{x}^K, \mathbf{y})$ depends on the size divergence when compared with others. In addition, we consider the improvement of generative diversity via the combination between the sampled hidden state $\mathbf{z}$ and random noise $\eta \sim \mathcal{N}(0, I)$ with the formulation as:

$$\hat{\mathbf{z}} = \mu_z(\tilde{\mathbf{x}}, \mathbf{y}) + \eta \otimes \sigma_z(\tilde{\mathbf{x}}, \mathbf{y}), \tag{5.2}$$

where $\tilde{\mathbf{x}} = \{\mathbf{x}^k\}_{k=1}^K$, $\mathbf{y}$ is the corresponding annotation and μ_z, σ_z are statistics of latent variable. Finally, the k-th source domain obtains enough reliable samples from $\hat{\mathbf{z}}$.

5.2.3 Generative Inference Network (GINet)

Basic Module. Along with the conventional DG methods [50], the basic module not only extracts domain-invariant features via the adversarial relationship between feature extractor and discriminator but also learn robust classifier with source supervisions. From Fig. 5.1, the framework involves feature extractor as ResNet [76] or AlexNet [146] which transforms input images into low-dimensional features, i.e., $\mathbf{x}_i = \mathbf{f}(\mathbf{X}_i)$ where $\mathbf{x}_i \in \mathbb{R}^d$ and $\mathbf{f}(\cdot)$ denotes the mapping function. Classifier then exploits these representations to predict the corresponding class probability distributions through $\mathbf{c}(\mathbf{x}_i)$. Due to the accessibility of annotation, it is simple to train the feature extractor and classifier by minimizing the following cross-entropy loss:

$$\mathcal{L}_c = -\frac{1}{N_s}\sum_{i=1}^{N_s}\mathbf{y}_i \log\big(\mathbf{c}\circ\mathbf{f}(\mathbf{X}_i)\big), \quad N_s = \sum_{k=1}^{K} N_k. \tag{5.3}$$

Moreover, the adversarial training manner is typically adopted to obtain domain-invariant representation on domain adaptation tasks [34]. Motivated by the application of GRL component [36], we develop a discriminator following $\mathbf{x}_i$ to identify which domain it belongs to. That is, each instance will automatically be attached with the domain-specific label $\ell_i \in \{1, 2, \cdots, K\}$. The parameter of discriminator $\mathbf{d}(\cdot)$ will be optimized with the minimization of cross-entropy loss as:

$$\mathcal{L}_d = -\frac{1}{N_s}\sum_{i=1}^{N_s}\ell_i \log\big(\mathbf{d}\circ\mathbf{f}(\mathbf{X}_i)\big). \tag{5.4}$$

The gradient derived from the above loss of discriminator is inversely propagated into the feature extractor to gradually eliminate domain-related attributions from hidden representations. Although such an adversarial learning strategy to some extent improves the generalization of model, it still difficultly receives sufficient knowledge from smaller source domains or categories to avoid the predictive partiality from the current learning system.

Balanced Generative Module. According to our proposed BGP, we instantiate it by developing balanced generative module to remedy the disadvantage of the existing DG models. Concretely, the module first builds two sub-networks $\mathbf{h}_\mu(\cdot)$ and $\mathbf{h}_\sigma(\cdot)$ to estimate $p(\mathbf{z}|\mathbf{x}^1, \cdots, \mathbf{x}^K, \mathbf{y})$ and deduce the statistics of latent variable $\mathbf{z}$, i.e., μ_z=$\mathbf{h}_\mu(\mathbf{x}, \mathbf{y})$ and σ_z=$\mathbf{h}_\sigma(\mathbf{x}, \mathbf{y})$, where $\mathbf{x} = \frac{1}{K}\sum_{k=1}^{K}\mathbf{x}^k$. It is worth nothing that the selected samples across different domains within this input share the identical annotation. Based on Eq. (5.2), we can sample sufficient latent variables $\hat{\mathbf{z}}$ to generate novel instances $\hat{\mathbf{x}}_i^k = \mathbf{g}^k(\hat{\mathbf{z}}_i)$, where $\mathbf{g}^k$ corresponds to the generator of the k-th source domain. To generate discriminative instances without domain-specific attribution, $\hat{\mathbf{x}}_i^k$ is also fed into the trained multi-class classifier $\mathbf{c}(\cdot)$ and multi-domain discriminator $\mathbf{d}(\cdot)$ to achieve the predictions as $\hat{\mathbf{y}}_i = \mathbf{c}\circ\mathbf{g}^k(\hat{\mathbf{z}}_i)$ and $\hat{\ell}_{\mathbf{i}} = \mathbf{d}\circ\mathbf{g}^k(\hat{\mathbf{z}}_i)$, respectively.

To this end, we need to introduce suitable objective function to optimize networks $\mathbf{h}_\mu(\cdot)$,

$\mathbf{h}_\sigma(\cdot)$ and $\mathbf{g}^k(\cdot)$ so that they can generate meaningful novel samples. Of course, we can adopt strict reconstruction loss as typical VAE methods [147]. However, such loss function heavily reduces the diversity of generative samples, which results in invalid generation for the balance of data scale. From another viewpoint, the original observable samples and novel generative instances forms two different distributions. And the reliable and meaningful novel instances also mean they should follow the similar even identical distribution with the original data. Hence, we expect to build the objective function from the perspective of distribution alignment. Motivated by the great success of optimal transport (OT) [148] on solving distribution alignment [34], we propose the objective function under OT framework to learn the optimal network parameters. For the convenient illustration, we take the generation procedure of one source domain as an example to introduce details by obtaining one novel instance $\hat{\mathbf{x}}_i$ for each instance $\mathbf{x}_i$. Specifically, $(\mathbf{x}_i, \mathbf{y}_i) \sim P$ and $(\hat{\mathbf{x}}_i, \hat{\mathbf{y}}_i) \sim \hat{P}$ are firstly considered as two pairs of random variables which are sampled from their corresponding complete metric space Ω and $\hat{\Omega}$ and the OT cost is defined as:

$$\begin{aligned} c: \big((\mathbf{x}_i, \mathbf{y}_i), (\hat{\mathbf{x}}_i, \hat{\mathbf{y}}_i)\big) \in \Omega \times \hat{\Omega} \\ \Rightarrow c\big((\mathbf{x}_i, \mathbf{y}_i), (\hat{\mathbf{x}}_i, \hat{\mathbf{y}}_i)\big) \in \mathbb{R}^+. \end{aligned} \tag{5.5}$$

Therefore, the loss function of minimizing distribution discrepancy is formulated as:

$$\mathcal{L}_g = \inf_{\pi} \mathbb{E}_{\big((\mathbf{x}_i, \mathbf{y}_i), (\hat{\mathbf{x}}_i, \hat{\mathbf{y}}_i)\big) \sim \pi} [c\big((\mathbf{x}_i, \mathbf{y}_i), (\hat{\mathbf{x}}_i, \hat{\mathbf{y}}_i)\big)], \tag{5.6}$$

where π is the distribution over $\Omega \times \hat{\Omega}$ and has marginals equal to P and $\hat{P}$. From our proposed balanced generative paradigm, it is simple to know that the latent causal can independently derive the cross-domain samples and the corresponding annotation. Based on this point, the OT cost function over the joint distribution of sample and label space can be divided into two individual cost functions for the convenient computation. With the theorem of [149], we can reformulate the above loss function as the following:

$$\begin{aligned} \mathcal{L}_g = \inf_{\mathbf{h}_\mu, \mathbf{h}_\sigma, \mathbf{g}} \mathbb{E}_{\mathbf{x}_i \sim P_x} \mathbb{E}_{\mathbf{y}_i \sim P_y} \mathbb{E}_{\hat{\mathbf{z}}_i \sim P(\mathbf{z}|\mathbf{x}, \mathbf{y})} [c_{\mathbf{x}}\big(\mathbf{x}_i, \mathbf{g}(\hat{\mathbf{z}}_i)\big) \\ + c_{\mathbf{y}}\big(\mathbf{y}_i, \mathbf{c} \circ \mathbf{g}(\hat{\mathbf{z}}_i)\big)], \end{aligned} \tag{5.7}$$

where P_x, P_y are the marginal distribution from P. When designing cost function of sample space, we consider that the generative features should be close to or similar with the original data points. Hence, we adopt Euclidean distance to measure and minimize their difference. On the other hand, we explicitly know that each novel sample is derived from which category latent variable. To preserve such category-wise semantics, we utilize the cross-entropy loss as the cost function. These considerations are embedded into the final objective function as:

$$\mathcal{L}_g = \frac{1}{N_s}\sum_{i=1}^{N_s}\Big(-\mathbf{y}_i \log\big(\hat{\mathbf{y}}_i\big) - \ell_i \log\big(\hat{\ell}_i\big) + \lambda\|\mathbf{x}_i - \hat{\mathbf{x}}_i\|_2^2\Big), \tag{5.8}$$

where λ is trade-off parameter between three loss terms.

Training and Inference. To this end, the above discussion has provided details of our proposed GINet including basic and balanced generative modules. To achieve stable training, we iteratively optimize these two modules by freezing one of them until convergence. The two sub-problems of our GINet model are formulated as follows:

$$\begin{cases} \min\limits_{\mathbf{f}(\cdot),\mathbf{c}(\cdot),\mathbf{d}(\cdot)} \mathcal{L}_c + \mathcal{L}_d, \\ \\ \min\limits_{\mathbf{h}_\mu(\cdot),\mathbf{h}_\sigma(\cdot),\mathbf{g}(\cdot)} \mathcal{L}_g, \end{cases} \tag{5.9}$$

where these two minimization optimizations are alternatively performed until convergence. For the inference stage, the unseen target data is fed into the basic module to obtain their prediction.

Moreover, in practical implementation, to guarantee the balance of data scale, we first randomly select several categories, and then collect the same number of images per category per domain to form one training batch. For example, for PACS dataset, each batch includes 60 samples uniformly distributed in three source domains. And each domain involves same 5 categories with 4 images per class. These original images are fed into the feature extractor to output their high-level features. Next, we divide the extracted feature into 5 groups with their class label. And then, for each group, we infer its statistics via network $\mathbf{h}_\mu$ and $\mathbf{h}_\sigma$ which takes the average of features from the same category and annotation as input. And then, we adopt reparameterization trick as Eq. 5.2 to sample enough latent variables to generate novel features for the corresponding class in the k-th source domain. Considering the balance of data scale, the generative procedure produces the same number of features for the selected category of each domain.

5.2.4 Theoretical Analysis

To illustrate the reliability of generative samples, we theoretically analyse the error bound about generative annotation and ground truth in Theorem. We first give one required Lemma and then use it to derive the Theorem 1.

Lemma. In probability theory, suppose random variable x comes from the sample space $\Omega = \{1, 2, \cdots, n\}$ with the corresponding probabilities $P_1 \leq P_2 \leq \cdots \leq P_n$, where $P_i = P(x = i) \geq 0, \sum_{i=1}^{n} P_i = 1$. Under this condition, we have the conclusion:

$$1 - P_n \leq 2(\sum_{i=1}^{n-2} \sum_{j=i+1}^{n-1} P_i P_j + \sum_{i=1}^{n-1} P_i P_n). \tag{5.10}$$

Theorem 1. Given the prior probabilities of multiple source domains $\{P_s^1, P_s^2, \cdots, P_s^K\}$ and the corresponding label probabilities within the k-th source domain $\{Q_k^1, Q_k^2, \cdots, Q_k^C\}$, and the probability densities of the latent variable $\{q_z^1, q_z^2, \cdots, q_z^C\}$ where $q_z^c(\mathbf{z}) = q(\mathbf{z}|\mathbf{y} = c)$, the error bound of the generative annotation is formulated as the following with the generalization error ϵ:

$$\begin{aligned} & |E(\mathbf{y}) - E(\hat{\mathbf{y}})| \\ & = 1 - \int \max \Big\{ \sum_{k=1}^{K} P_s^k Q_k^1 q_z^1, \cdots, \sum_{k=1}^{K} P_s^k Q_k^C q_z^C \Big\} d\mathbf{z} \leq \epsilon, \end{aligned} \tag{5.11}$$

where $\hat{\mathbf{y}}$ is the generative annotation from classifier.

Remark. Theorem 1 suggests that our generative manner not only augments more training samples to generalize classifier but also sufficiently preserves semantic information related to classification task in the synthesized instances. Thus, our method effectively extends original data with high-quality and reliable novel samples to reduce the negative effect of imbalanced data distribution across various domains and categories. The specific

Table 5.1: Comparisons of Object Recognition Rate (%) for Domain Generalization task on PACS benchmark under **Normal** setting. The best performance is highlighted in **bold**, while the second highest result is shown with underline. (Backbone: AlexNet)

PACS	$\mathcal{D}_t$	CIDDG	MetaReg	MASF	Epi-FCR	JiGen	DMG	EISNet	Ours
AlexNet	**P**	78.65	91.07	90.68	86.1	89.00	87.31	91.20±0.00	**91.7±0.1**
	A	62.70	69.82	70.35	64.7	67.63	64.65	70.38±0.37	**73.0±0.4**
	C	69.73	70.35	**72.46**	72.3	71.71	69.88	71.59±1.32	72.1±0.9
	S	64.45	59.26	67.33	65.0	65.18	**71.42**	70.25±1.36	70.9±1.1
Average		68.88	72.62	75.21	72.0	73.38	73.32	75.86	**76.9**

proof is in Appendix[1].

5.3 Experiments

5.3.1 Experimental Setup

Benchmarks. 1) PACS [51] is the recent widely-used domain generalization benchmark for object recognition including 9,991 visual signals of seven categories shared by four domains: **P**hoto, **A**rt Painting, **C**artoon and **S**ketch with considerable domain shift. **2) VLCS** [150] as the classic benchmark for DG task is composed of four domains drawn from PASCAL **V**oc2007, **L**ableme, **C**altech-101 and **S**un09 with images distributed in five categories. The specific data distribution across various domains and categories over PACS and VLCS are reported in Appendix[1]. With respect to them, these benchmarks both contain significant difference of data scale across various domains and categories. **3) Office-Home** includes four domains, i.e., Realworld (**Rw**), Clipart (**Cl**), Art (**Ar**), Product (**Pr**) with each domain from 65 categories. The specific sample size for each domain is **Ar** (2,427), **Cl** (4,365), **Pr** (4,439) and **Rw** (4,357), respectively. Following the same protocol in [145], any three domains per benchmark are used as multiple source sets while the left one serves as target domain.

The original benchmarks are somehow imbalanced. We evaluate the original data by defining **Normal Setting**, and further considering **Imbalanced Setting**. Specifically, for the considered imbalanced domain generalization on the mentioned datasets, we select one of

[1] https://github.com/HaifengXia/IDG/appendix.pdf

Table 5.2: Comparisons of Object Recognition Rate (%) for Imbalanced Domain Generalization task on PACS benchmark. The best performance and the second one are highlighted in **bold** and underline. (Backbone: AlexNet, ResNet-18 and ResNet-50)

PACS	$\mathcal{D}_t$	Epi-FCR [52]	JiGen [54]	DMG [145]	EISNet [53]	RSC [139]	FACT [49]	Ours
AlexNet	**P**	78.36±0.68	82.68±0.42	77.67±0.55	86.18±0.22	83.75±0.29	84.34±0.42	**86.49±0.29**
	A	58.85±0.49	60.97±0.47	59.35±0.41	62.43±0.32	60.22±0.11	64.05±0.23	**67.15±0.27**
	C	69.05±0.48	68.65±0.34	66.84±0.35	69.53±0.22	69.93±0.18	**70.62**±0.31	69.75±0.26
	S	59.33±0.54	59.24±0.38	64.69±0.46	65.12±0.28	66.94±0.23	67.13±0.32	**68.53±0.14**
Average		66.3	67.8	67.0	70.7	70.2	71.5	**73.0**
ResNet-18	**P**	92.17±0.42	92.85±0.37	92.28±0.48	94.03±0.24	92.56±0.25	93.27±0.36	**96.05±0.23**
	A	70.73±0.37	74.33±0.32	71.35±0.62	76.94±0.34	74.63±0.35	78.41±0.27	**79.26±0.18**
	C	64.76±0.31	71.25±0.33	**74.56±0.22**	70.83±0.29	72.26±0.45	72.58±0.30	74.46±0.19
	S	63.92±0.46	65.35±0.26	68.26±0.38	68.78±0.29	69.27±0.48	**70.67±0.13**	70.64±0.28
Average		72.9	75.9	76.6	77.6	77.1	78.7	**80.1**
ResNet-50	**P**	94.66±0.40	96.37±0.35	92.83±0.38	97.02±0.26	94.92±0.27	95.36±0.31	**98.09±0.24**
	A	80.27±0.48	79.66±0.32	78.05±0.39	82.58±0.25	81.33±0.27	83.51±0.24	**85.53**±0.13
	C	76.42±0.33	74.84±0.24	76.63±0.19	76.79±0.38	76.44±0.27	76.95±0.24	**76.98±0.22**
	S	74.04±0.26	73.28±0.31	76.73±0.24	76.14±0.17	77.23±0.21	**78.15±0.12**	77.96±0.29
Average		81.3	80.9	81.0	83.1	82.5	83.4	**84.6**

four from the original dataset as the target domain without any change. The remaining ones are considered as source domains where we randomly remove instances across domain and category to produce the divergence of data scale. There are four domain generalization tasks within each benchmark. The details of imbalanced scenario are summarized in Appendix[1].

Baselines. We compare our GINet with several state-of-the-art deep domain generalization methods including meta-learning solutions, i.e., MetaReg [151] and MASF [142], generative adversarial strategy, i.e., CIDDG [50], jigsaw puzzle auxiliary task based on solutions, i.e., JiGen [54] and EISNet [53], episodic training based on approach Epi-FCR [52], other methods such as Mixup [152], FACT [49], DMG [145], RSC [139], DAEL [138], DIFEX [153] and Vrex [154]. We re-implement the above methods with their public available code and perform the optimal parameter selection to make a fair comparison with our method on

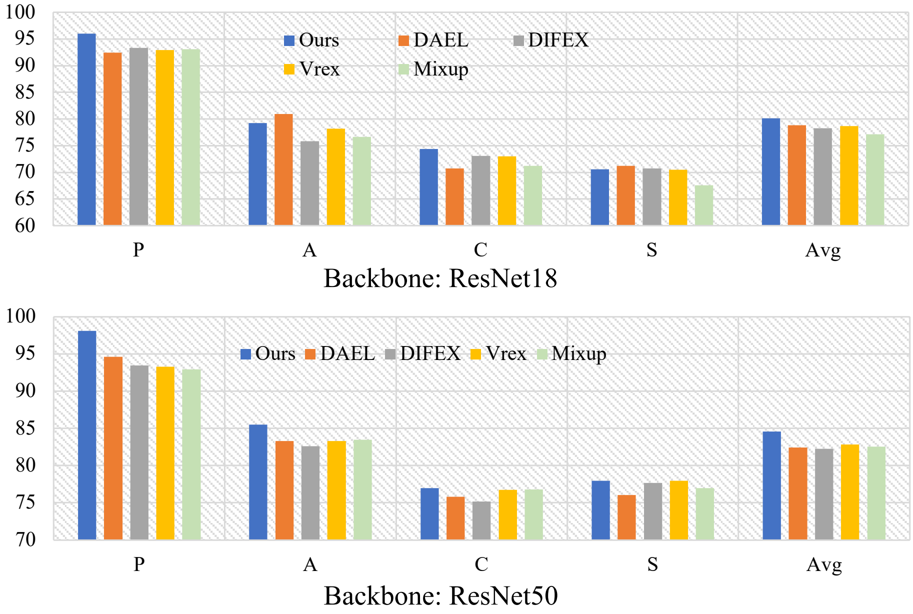


Figure 5.3: Comparison of classification accuracy between our method and Mixup [152], DAEL [138], Vrex [154] and DIFEX [153] under IDG scenarios on PACS dataset with different backbones.

both normal and imbalanced DG settings.

Implementation Setting. In terms of the feature extractor, we follow others [52] and consider various backbones like AlexNet [146], ResNet-18 [76], ResNet-50 [76] pre-trained on ImageNet [11] without the last layer. The classifier includes one full-connected (FC) layer with the same number of input as the previous feature and the same number of output as category number (C). And the discriminator includes three FC layers, i.e., 1024→256→K. For the balanced generative module, we firstly use one FC layer (1024→64) to calculate mean and co-variance and then adopt another three FC layers (64 →512→1024) to build the generator with ReLU and batch normalization. As the optimizer, we train the network with stochastic gradient descent (SGD) with momentum 0.9. Moreover, the learning rate γ_t is adjusted by $\gamma_t = \dfrac{\gamma_0}{(1+\alpha t)^b}$ where $\alpha = 10$, $b = 0.75$ and t is linearly increase from 0 to 1. The initial γ_0 is set as 0.001 for feature extractor and 0.01 for other components. In terms of the selection of hyper-parameter λ, we determine it with the performance of model on validation set. Concretely, for each task, we randomly select 90% source images for training and consider the remaining ones as validation set. Given the specific λ, we can obtain the well-trained model and evaluate it on validation set. Next, we deploy the optimal model on the unseen target domain and this model achieves the highest classification accuracy on

Table 5.3: Comparisons of Object Recognition Rate (%) for DG task on PACS benchmark under **Normal** setting. The best performance is highlighted in **bold**, while the second highest result in shown with underline. (Backbone: ResNet-18 and ResNet-50).

PACS	Method	**P**	**A**	**C**	**S**	Avg.
ResNet-18	Mixup [152]	95.2	81.8	75.4	76.5	82.2
	Epi-FCR [52]	93.9	82.1	77.0	73.0	81.5
	JiGen [54]	96.0	79.4	75.3	71.4	80.5
	MASF [142]	95.0	80.3	77.2	71.7	81.5
	DMG [145]	93.4	76.9	**80.4**	75.2	81.5
	EISNet [53]	95.9	81.9	76.4	74.3	82.2
	RSC [139]	94.4	80.5	78.6	76.0	82.3
	FACT [49]	94.8	83.7	78.3	77.3	83.5
	DAEL [138]	95.6	**84.6**	74.4	78.9	83.4
	Vrex [154]	95.3	81.5	77.5	78.1	83.5
	DIFEX [153]	95.5	80.8	77.6	**79.4**	83.3
	Ours	**96.8**	84.2	78.9	75.7	**83.7**
ResNet-50	Mixup [152]	96.6	87.3	80.6	82.6	86.7
	Epi-FCR [52]	95.8	84.9	81.3	76.6	84.6
	JiGen [54]	96.9	84.5	80.4	75.5	84.3
	MASF [142]	94.5	82.9	80.5	72.3	82.7
	DMG [145]	94.5	82.6	78.1	78.3	83.4
	EISNet [53]	97.1	86.6	**81.5**	78.1	85.8
	RSC [139]	95.1	83.9	79.5	82.2	85.1
	FACT [49]	95.5	87.2	80.9	**83.6**	86.8
	DAEL [138]	96.6	86.8	80.4	81.7	86.4
	Vrex [154]	96.9	87.1	**81.5**	82.8	87.0
	DIFEX [153]	96.4	86.6	80.1	82.3	86.4
	Ours	**98.0**	**89.0**	**81.5**	80.2	**87.2**

validation set.

5.3.2 Comparison Results

Table 5.1 and Table 5.3 summarize the object recognition accuracy on four domain generalization tasks under normal setting with the original benchmark PACS. Due to the prolific semantic knowledge captured by deeper network architecture, all domain generalization methods based on ResNet-18/50 yield significant improvement over AlexNet based. Averagely, our GINet outperforms other competitive baselines with three various backbones, achieving the state-of-the-art performance and highly affirming the effectiveness of

Table 5.4: Comparisons of Object Recognition Rate (%) for DG task on Office-Home under **Normal** and **Imbalanced** settings. The best performance is highlighted in **bold**, while the second highest result is shown with underline. (Backbone: ResNet-18).

Setting	Method	**Ar**	**Cl**	**Pr**	**Rw**	Avg.
Normal	Mixup [152]	58.7	51.0	72.2	75.4	64.3
	DSON [155]	59.4	45.7	71.8	74.7	62.9
	RSC [139]	57.6	48.4	72.6	74.1	63.1
	L2A-OT [115]	60.6	50.1	74.8	77.0	65.6
	FACT [49]	60.3	54.8	74.4	76.5	66.5
	DAEL [138]	59.4	**55.1**	74.0	75.7	66.1
	Vrex [154]	59.0	49.8	71.6	74.8	63.8
	DIFEX [153]	59.3	50.2	71.2	75.2	64.0
	Ours	**61.9**	52.7	**75.3**	**77.5**	**66.9**
Imbalanced	Mixup [152]	55.2	48.2	69.9	72.5	61.4
	DMG [145]	51.6	42.6	68.7	71.1	58.5
	EISNet [53]	52.8	44.5	70.4	73.7	60.3
	RSC [139]	55.1	43.4	69.5	71.9	60.0
	FACT [49]	56.2	**49.8**	71.3	73.6	62.7
	DAEL [138]	57.0	49.0	71.7	73.5	62.8
	Vrex [154]	56.1	47.3	69.1	72.3	61.2
	DIFEX [153]	55.4	47.5	68.4	73.1	61.1
	Ours	**59.7**	49.4	**73.7**	**75.7**	**64.6**

our method on solving domain generalization. Specifically, our method boosts the accuracy of CIDDG based on adversarial learning strategy by a absolute 8.0% when adopting AlexNet. This suggests that our GINet effectively facilitates the generalization of model through instance augmentation. When compared with EISNet producing the second highest average accuracy, the promotion of GINet mainly derives from more accurate object recognition on unseen target domains **A** and **S**. With respect to the results with AlexNet, the classification accuracy of our method (73.0%) surpasses EISNet (70.4%) by 2.6% on **A** domain. The reason for this situation comes from that GINet generates more samples for the smallest source domain **P**, where EISNet difficultly learns sufficient knowledge, to improve the performance of classifier.

Moreover, we also evaluate the mentioned methods on imbalanced domain generalization scenario, where the divergence of data scale across various source domain and categories is further exacerbated than the original PACS. Table 5.2 reports the corresponding classification accuracy. Compared with Table 5.1 and Table 5.3, all baseline methods suffer from dramatic performance degradation with insufficient training samples in several domains or categories. However, according to the results, it is worth nothing that the advantage of our GINet over others becomes more conspicuous. For instance, our method with ResNet-18 (80.1%) outperforms EISNet (77.6%) by 2.5% on the average accuracy. This illustrates that our method exploits latent variable to capture complicated relationship among multiple source domains and synthesize reliable features to make classifier generalized.

Similarly, Table 5.5 and Table 5.4 report the performance of our GINet with that of other recent state-of-the-art methods under normal and imbalanced settings on VLCS and Office-Home benchmarks. According to the statistics of VLCS benchmark in Appendix[1], such two scenarios have significant difference of data scale across various domains and categories. Different from PACS, there is the small cross-domain discrepancy in VLCS only with photo style. All solutions, thus, show the stable recognition ability under two experimental settings. But the diversity of training sample still is important for the improvement of generalization, which is verified by the higher classification accuracy of GINet and EISNet than DMG. Although EISNet utilizes jigsaw puzzle manner to learn inherent features and augment instances, the learned direct cause of multi-source domains facilitates our method to achieve better generalization, specifically on **V** task (79.5% vs 74.8%). In addition, learning a generalized model becomes more challenging on Office-Home benchmark since it includes more object categories than PACS and VLCS. However, our proposed method still achieves comparable even better accuracy than others in most tasks. Especially, with **Ar** as the unseen target domain, our GINet fights off the second best result (FACT) by a larger margin for IDG scenario. It is worth nothing that FACT actually augments more images by using Fourier transformation fashion, which learns domain-invariant representations to generalize the model on the unseen target domain yet difficultly addresses the bottleneck of imbalanced sample distribution across different domains and categories. The comparison between GINet and FACT further illustrates that our balanced generative strategy effectively mitigates the negative influence of imbalanced data distribution on learning generalized classification model for unseen target domain.

Table 5.5: Comparisons of Object Recognition Rate (%) for DG task on VLCS under **Normal (N)** and **Imbalanced (I)** settings. The best performance is highlighted in **bold**, while the second highest result is shown with underline. (Backbone: ResNet-50).

VLCS	Method	**V**	**L**	**C**	**S**	Avg.
Normal	Mixup [152]	75.0	66.2	96.9	68.4	76.6
	DMG [145]	73.7	**68.3**	97.0	70.7	77.4
	EISNet [53]	75.6	66.9	97.6	71.0	77.8
	RSC [139]	75.8	66.1	96.9	70.3	77.3
	FACT [49]	75.9	66.3	96.9	**71.4**	77.6
	DEAL [138]	75.0	66.2	96.9	69.3	76.9
	Vrex [154]	76.9	66.1	95.2	**71.4**	77.4
	Ours	**79.7**	65.6	**98.6**	71.2	**78.8**
Imbalanced	Mixup [152]	72.0	63.2	96.7	67.6	74.9
	DMG [145]	73.1	65.3	96.3	69.3	76.0
	EISNet [53]	74.8	65.3	97.0	69.5	76.7
	RSC [139]	75.6	**65.6**	97.2	69.7	77.0
	FACT [49]	75.6	64.8	96.0	**71.0**	76.9
	DEAL [138]	73.6	62.8	94.6	68.4	74.9
	Vrex [154]	74.5	65.2	95.5	69.2	76.1
	Ours	**79.5**	64.5	**98.1**	70.4	**78.1**

5.3.3 Empirical Analysis

Feature Visualization. To analyse the effect of imbalanced training data on learning a generalized classifier, we specifically show the visualization of target features derived from ResNet-50 with **V** as the unseen target domain in Figure 5.5. Since the training sample size of facial images (purple) is many times larger than dog photo (blue) in multi-source domains. Thus, the baseline models learning more knowledge from larger categories fail to acquire discriminative information from others. Oppositely, our GINet with balanced generative paradigm reduces the negative influence of the larger category or domain on learning generalized features and produces clear classification boundary in Figure 5.5-(c).

Ablation Study & Parameter Analysis. To clearly reflect the effect of our proposed balanced generative paradigm (BGP), Figure 5.6-(a) reports the comparison between GINet variant without BGP module and the overall network architecture on imbalanced VLCS benchmarks. According to the results, GINet surpasses the variant by a large margin for

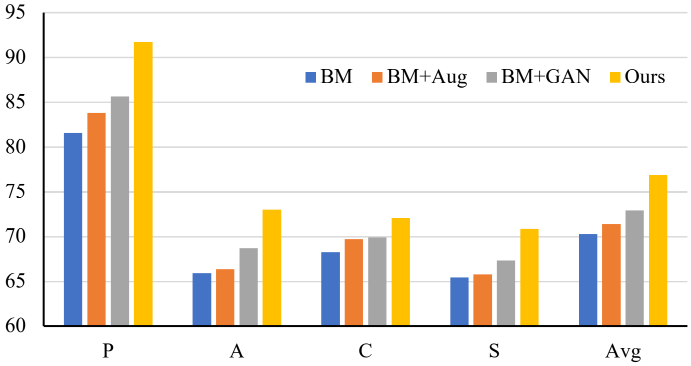
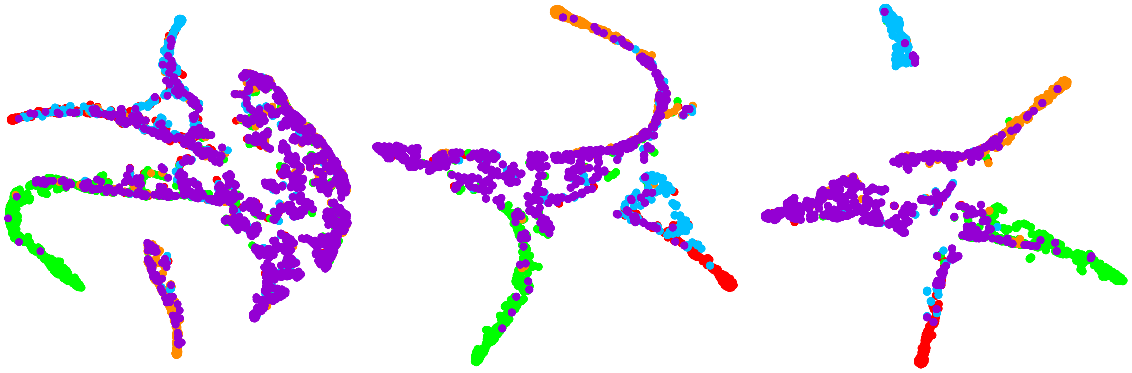


Figure 5.4: Comparison of classification accuracy between our method and many data augmentation techniques.



(a) DMG

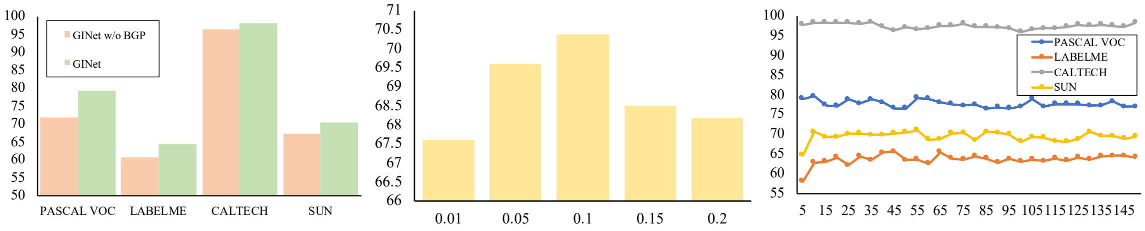
(b) EISNet

(c) Ours (GINet)

Figure 5.5: t-SNE embedding visualization of three different models, where five category-color pairs are listed as bird-red, car-green, chair-orange, dog-blue and person-purple.

four IDG tasks, particularly, when evaluated on Pascal-**V**oc domain. It demonstrates the BGP makes more contributions on enhancing generalization, since this module effectively captures intrinsic features from multi-source domains to further facilitate the performance of classifier.

On the other hand, the performance improvement obtained by our method results from the combination data augmentation and the balance of data scale. To analyze which part makes the main contribution, we design the corresponding experiments to illustrate this point. Concretely, we adopt random sampling manner over source domains to build each training batch and only use the basic module optimized by Eq. 5.3 and 5.4, which is named as BM. Based on BM, we utilize some normal image transformation such as contrast adjustment, horizontal flip, etc, to augment training samples and name it as BM-Aug. In addition, we consider using conditional generative adversarial network (GAN) to conduct data aug-



(a) Variant Evaluation

(b) Parameter λ

(c) Number of Epoch

Figure 5.6: Performance analysis including ablation study, parameter selection and training stability.

mentation. Specifically, we introduce additional generator and discriminator into BM where the generator takes the combination of domain label, class label and random noise as input to generate fake hidden features, which is denoted as BM-GAN. In addition, BM, BM-GAN and our method use the same number of iterations, while the training iteration of BM-Aug is two times than others. Note that the total number of generative and original samples of BM-GAN, BM-Aug and ours is two times than that of BM. The experiments are conducted on the original PACS dataset via AlexNet with results in Fig. 5.4. From the comparison, we can find that the simple data augmentation strategies as BM+Aug and BM+GAN indeed result in positive effect on performance improvement. Especially, BM+GAN advances the baseline BM about 2.6% w.r.t the average of all tasks. However, our method (GINet) outperforms BM and BM+GAN by 6.7% and 4.0%, respectively. Therefore, we can have the conclusion that the performance improvement of our proposed method mainly comes from the increasing number of training sample and the balance of data scale across domains and categories. And the positive contribution of the balance of data scale is much larger than that of data augmentation. For the hyper-parameter λ in our method, we select it from $\{0.01, 0.05, 0.1, 0.15, 0.2\}$ and record an example performed on unseen **Sun** domain with imbalanced VLCS dataset. As Fig. 5.6 (b) shows, the adjustment of λ directly affects the quality of generative instances and model performance.

Training Stability. The iterative training manner is adopted to alternatively update basic and balanced generative modules, which might affect the stability of model. Therefore, we investigate the relationship between classification accuracy on target domain and the training process (epoch) in Figure 5.6-(c). From the results on imbalanced VLCS, we notice that there are small fluctuations on accuracy with the increasing number of epoch. But the overall process has no considerable performance degradation, which suggests such a training strategy is reasonable and reliable.

5.4 Conclusions

Domain Generalization (DG) attempts to learn a generalized model with multi-source domains and directly adapt it into the unseen target domain. In this paper, we empirically discovered that the imbalanced data scale across various source domains and categories would negatively affect the learning of robust transferable representations. Thus, we formulated it as imbalanced domain generalization (IDG) scenario and presented a novel model as Generative Inference Network (GINet). The core idea of GINet was to explore the accessible source instances to infer their direct cause which uniformly reflects the attributions of each domain and category. With the learned latent variable, GINet introduced a Balanced Generative Module to augment more reliable features to facilitate the generalizability of classifier. Theoretical analysis guarantees the quality of augmented samples and extensive empirical studies on three popular benchmarks verified the advantage of GINet over the state-of-the-art DG methods.

Chapter 6

Representation Decomposition for Zero-Shot Domain Adaptation

6.1 Background

Computer vision community always suffers from insufficient annotation issue, which dramatically obstructs the practical applications of most techniques. However, domain adaptation provides an alternative strategy to handle with such a problem [59, 62, 82]. Concretely, it attempts to borrow knowledge from well-annotated modality (source domain) to solve classification task on target domain without any label information [22, 67, 96]. Although various domains share the high-level semantic information, their data distributions contain significant discrepancy defined as domain shift [55, 66, 69]. For example, due to light condition or occlusions, visual instances involving the same object are different from each other [56]. As a result, the previously-trained model generally tends to be fragile when evaluated on target domain.

Domain adaptation (DA) as a solution to learn domain-invariant knowledge attracts great interest [63, 68, 84, 120]. To learn transferable information, it assumes that instances of target modality are available [65, 74, 81]. Under such an assumption, recent works mainly explore two approaches: discrepancy measurement [27] and domain adversarial confusion [22, 64]. Specifically, the first strategy aims to define novel statistic indicators like maximum mean discrepancy (MMD) [63] promoting the consistency of distribution. While methods based on domain adversarial confusion expect to transform data of source and target domain into the similar hidden space by using adversarial relationship between generator and discriminator. They actually have achieved promising improvement in distinctive tasks. In real-world scenarios, however, the assumption which they depend on is infeasible due to the absence of target domain. The general situation is defined as *zero-shot domain adaptation* (ZSDA) [156], which is also known as *missing modality transfer learning* [157]. For instance, to protect privacy of patient, hospital fails to share medical records to train the model, even though they expect to apply the trained model for their work, where these documents represent target domain. In this sense, the current DA methods are more likely to be invalid since the guidance of target datatset becomes invisible.

The awkward situation inspires [158] to proposes domain-invariant component analysis (DICA) by using multiple source domains with identical label space to build a generalized model for unseen target recognition. However, they hardly collect sufficient source domains to observe the information of unseen target modality. To solve this problem, the intuitive motivation is to introduce auxiliary task-irrelevant dataset (TIR), which also includes two same modalities with the task-relevant one (TR) [157]. Alternatively, [156] develops the first deep model for zero-shot domain adaptation which firstly attempts to achieve the feature alignment on task-irrelevant datasets and then allows source modalities in TR and TIR to share the same network. Moreover, the generalization of neural network facilitates the consistency of cross-domain distribution on task-relevant dataset. Albeit the training manner enables model to generate domain-invariant representation, features tend to be less discriminative without the guidance of annotation when training model on task-irrelevant inputs, leading to the decrease of recognition. Meanwhile, due to the huge achievement of generative adversarial model in abundant practical scenarios, it is appropriate to utilize this manner to synthesis missing modality and directly perform domain adaptation in TR datasets [159] named CocoGAN. However, the drawbacks of generative adversarial network is that there exists bias between generated instances and real samples, since synthesised images only try to approximate the real distribution. Thus, estimating the influence of bias on the final classification task tend to be very difficult. On the other hand, we naturally post a question about CocoGAN: “Is the explicit generation of missing target dataset necessary for learning domain-invariant feature?”.

To answer this question, we rethink Zero-shot Domain Adaptation from feature separation and propose Hybrid Generative Network (HGNet), which not only synthesises domain-invariant feature but also effectively facilitates high-level representation to be more discriminative. Specifically, the whole network architecture mainly consists of four components: feature extractor, adaptive feature separation module, hybrid generator and classifier. Input signals of TR and TIR datasets firstly pass through feature extractor and are transformed into shallow convolutional units. For the second step, feature separation module adaptively selects several channels to form classification-relevant high-level feature, while others are considered as classification-irrelevant information. In the final stage, on one hand, we apply the supervision of annotation to learn more discriminative units. On the other hand, hybrid generator will integrate object context and domain information belonging to various datasets to reconstruct input data. Extensive experimental performances illustrate that the hybrid strategy guarantees the uniqueness of feature separation as well as the completeness of semantic information. The contributions of our method are summarized in three folds:

- From the perspective of feature separation, we introduce a novel strategy named Hybrid Generative Network (HGNet) to fight off ZSDA more effectively. The proposed feature separation module guided by annotation explores global information from shallow convolutional layers to extract more discriminative and domain-invariant units.

- To perform high-quality feature separation, we develop hybrid generation module assisting model to capture association between task-relevant (TR) and task-irrelevant (TIR) datasets. The benefit of such a relationship is to utilize cross-domain knowledge learned from TIR to eliminate domain shift on TR datasets.

- We assess our model on several visual cross-domain tasks, and HGNet outperforms competitive approaches by large margin in most cases, illustrating the effectiveness on solving ZSDA challenge. We further conduct extensive empirical study to demonstrate the function of hybrid generation.

6.2 The Proposed Method

6.2.1 Preliminaries and Motivation

Zero-shot Domain Adaptation aims to exploit all accessible data to learn robust and generalized model used to deal with classification issue on target domain. Concretely, we are given well-annotated task-relevant source dataset $\mathcal{D}^{r,s} = \{(\mathbf{X}_i^{r,s}, Y_i^{r,s})\}_{i=1}^{n}$, where $\mathbf{X}_i^{r,s}$ and $Y_i^{r,s}$ separately denote i-th visual instance and its corresponding label. In addition, we also have access to task-irrelevant cross-domain paired datasets $\mathcal{D}^{ir,s} = \{(\mathbf{X}_i^{ir,s}, Y_i^{ir,s})\}_{i=1}^{m}$ and $\mathcal{D}^{ir,t} = \{(\mathbf{X}_i^{ir,t}, Y_i^{ir,t}\}_{i=1}^{m}$. Although $\mathbf{X}_i^{ir,s}$ and $\mathbf{X}_i^{ir,t}$ lie in various domains (source and target), they belong to the same category i.e., $Y_i^{ir,s} = Y_i^{ir,t}$. To this end, it is impossible for model to capture any knowledge of task-relevant target dataset $\mathcal{D}^{r,t} = \{\mathbf{X}_i^{r,t}\}_{i=1}^{n}$ only available in the test stage. The current scenario mainly involves two challenges: 1) **Generation of domain-invariant representation:** The absence of $\mathcal{D}^{r,t}$ results in huge difficulty of directly measuring cross-domain discrepancy between $\mathcal{D}^{r,s}$ and $\mathcal{D}^{r,t}$; 2) **Fusion of various datasets:** Tremendous difference among $\mathcal{D}^{r,s}$, $\mathcal{D}^{ir,s}$ and $\mathcal{D}^{ir,t}$ dramatically interferes their connection.

To capture domain shift between $\mathcal{D}^{r,s}$ and $\mathcal{D}^{r,t}$, the intuitive idea [159] is to firstly synthesize missing modality $\mathcal{D}^{r,t}$ and then transform them into the similar latent space, which arises a question: "*Is the explicit generation of missing target dataset necessary for learning domain-invariant feature*?" To answer this question, we rethink and explore the extraction of domain-invariant representation from the perspective of feature separation. Specifically, the intrinsic knowledge of input data generally is stored in high-level semantic representation via feature extractor. However, these semantic information is not equally necessary in terms of classification task. Admittedly, partial abstract representations record abundant essential content as visual style or background in object image, but they are drastically various across domains. We consider these representations as classification-irrelevant features, which are undesirable in domain adaptation. On the other hand, the remaining part defined as classification-relevant feature has positive influence on our final object classification task. Considering the previous approaches about domain-invariant feature learning, it is irrational or even counterproductive to incorporate all information into the same represen-

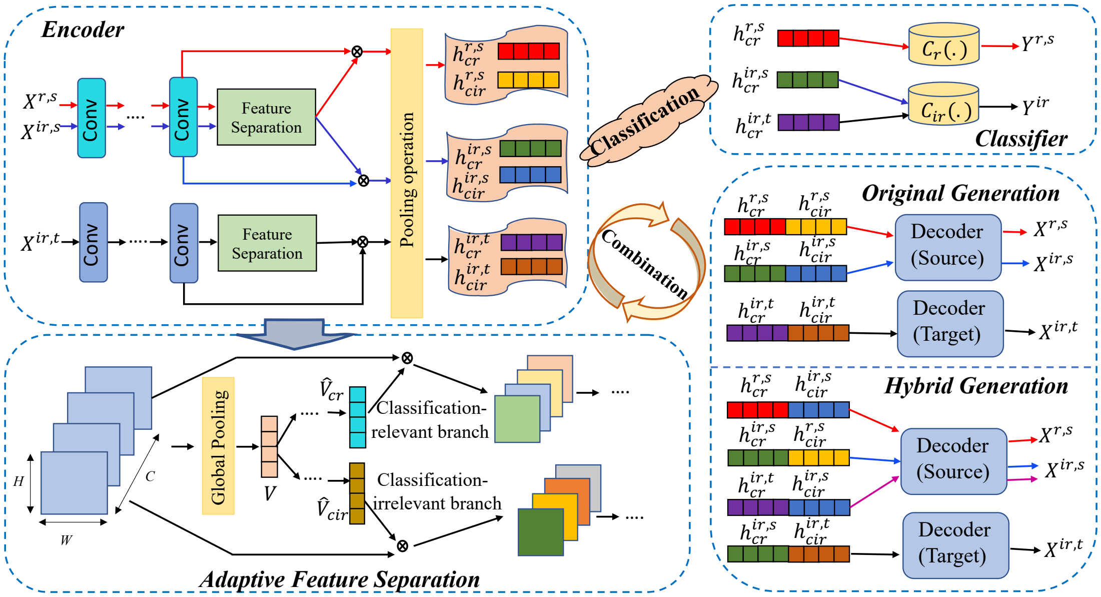


Figure 6.1: Overview of the proposed HGNet, which mainly includes four components: encoder, decoder, classifier and adaptive feature separation module. The encoder firstly aims to extract convolutional features, and then the adaptive feature separation module attempts to learn classification-relevant and classification-irrelevant units. On one hand, we utilize label information to guarantee the effect of feature separation. On the other hand, we explore two reconstruction manners to promote the completeness of semantic information and the uniqueness of learned feature.

tation. Therefore, we achieve two primary conclusions: 1) Feature separation is important to distinguish domain-invariant feature out of classification-irrelevant features instead of generating missing dataset $\mathcal{D}^{r,t}$; and 2) we should only explore discriminative information on the the selected classification-relevant representations. According to these discussions, we propose our adaptive feature separation module embedded into auto-encoder framework.

Due to feature separation, classification-irrelevant representations of instances from $\mathcal{D}^{r,s}$ and $\mathcal{D}^{ir,s}$ should preserve high-similarity. Such relationship is also applied to $\mathcal{D}^{r,t}$ and $\mathcal{D}^{ir,t}$. Cross-domain paired datasets $\mathcal{D}^{ir,s}$ and $\mathcal{D}^{ir,t}$ tend to be transformed into the same hidden space with respect to classification-relevant feature, which is also suitable for $\mathcal{D}^{r,s}$ and $\mathcal{D}^{r,t}$. Based on these above analyses, we develop hybrid reconstruction strategy to build the connection among various datasets and promote the performance of domain adaptation.

6.2.2 Adaptive Feature Separation

To effectively learn domain-invariant hidden units, we propose adaptive feature separation module, which is capable of distinguishing classification-relevant features from classification-irrelevant ones. As a result, the mechanism tends to describe same instance from two completely distinctive semantic views. To be specific, a branch of this module guided by discriminative information (annotation) aims to generate classification-relevant features, while the other branch will store other semantic contents. Moreover, auto-encoder framework combines them to reconstruct the input signal, which indeed guarantees the completeness of information and the difference between these two types of feature. From this property, we explore automatic feature selection from channel level.

Additionally, due to the generalization of deep neural network on feature learning, $\mathcal{D}^{r,s}$ and $\mathcal{D}^{ir,s}$ belonging to the same modality should share the network architecture and corresponding parameters. For $\mathcal{D}^{ir,t}$, the difference between source and target domain inspires us to adopt a distinctive network framework sharing parameters in higher network layers with the network for source domain. As shown in Figure 6.1, two various encoders involving convolutional operation convert the input signals $\mathbf{X}^{r,s}$, $\mathbf{X}^{ir,s}$ and $\mathbf{X}^{ir,t}$ into abstract representations $\mathbf{F}^{r,s}$, $\mathbf{F}^{ir,s}$, $\mathbf{F}^{ir,t} \in \mathbf{R}^{W \times H \times C}$, where W, H separately denote the width and height of each tensor, and C is the number of channel in tensor. At this time, the extracted features incorporate all semantic information of input data.

To learn domain-invariant features, we implement convolutional transformation to generate classification-relevant feature $\mathbf{F} \rightarrow \hat{\mathbf{F}}_{cr} \in \mathbf{R}^{W \times H \times C}$ and classification-irrelevant one $\mathbf{F} \rightarrow \hat{\mathbf{F}}_{cir} \in \mathbf{R}^{W \times H \times C}$, where $\mathbf{F}$ is selected from $\{\mathbf{F}^{r,s}, \mathbf{F}^{ir,s}, \mathbf{F}^{ir,t}\}$. The first transformation $\mathbf{F} \rightarrow \hat{\mathbf{F}}_{cr}$ performs a positive activation on convolutional layer via the guidance of label information to capture more discriminative information while gradually eliminating classification-irrelevant semantic content preserved in $\hat{\mathbf{F}}_{cir}$ with negative activation. Concretely, we firstly operate global average pooling technique on shallow convolutional feature $\mathbf{F}$ to obtain the information increment of each channel defined by $\mathcal{V} \in \mathbf{R}^{1 \times 1 \times C}$. Intuitively, each element $v_i \in \mathcal{V}$ roughly reflects content and style of the corresponding channel. To observe the connection across channels and separate features, we first adopt two distinctive non-linear manners to compress $\mathcal{V}$ to $\widetilde{\mathcal{V}}_{cr}$ and $\widetilde{\mathcal{V}}_{cir} \in \mathbf{R}^{1 \times 1 \times \frac{C}{\gamma}}$, where γ is a ratio controlling the scale of dimension-reduction and then utilize various full-connection layers to obtain new channel-wise statistics $\hat{\mathcal{V}}_{cr}$ and $\hat{\mathcal{V}}_{cir} \in \mathbf{R}^{1 \times 1 \times C}$. After the activation operation, $\hat{\mathcal{V}}_{cr}$ ideally promotes performance of several channels recording extensive discriminative information, while $\hat{\mathcal{V}}_{cir}$ enhances representation of others. Based on the above explanation, convolutional conversion can be formulated as:

$$\hat{\mathcal{V}}_{cr} = \sigma\Big(\mathbf{W}_{cr}\delta\big(g_{cr}(\mathcal{V})\big)\Big), \qquad \hat{\mathcal{V}}_{cir} = \sigma\Big(\mathbf{W}_{cir}\delta\big(g_{cir}(\mathcal{V})\big)\Big), \tag{6.1}$$

where $\mathbf{W}_{cr}$, $\mathbf{W}_{cir} \in \mathbf{R}^{C\times\frac{C}{\gamma}}$, $\sigma(\cdot)$ and $\delta(\cdot)$ represent **Sigmoid** and **ReLU** activation functions, $g_{cr}(\cdot)$ and $g_{cir}(\cdot)$ refer to the non-linear dimension-reduction operations. To achieve the feature separation based on classification-task, we conduct channel-wise multiplication ($\otimes$) between original convolutional features $\mathbf{F}$ and learned channel-wise indicators $\hat{\mathcal{V}}_{cr}$, $\hat{\mathcal{V}}_{cir}$ as the following:

$$\hat{\mathbf{F}}_{cr} = \hat{\mathcal{V}}_{cr} \otimes \mathbf{F} = \{\hat{v}_{cr,i} \cdot \mathbf{F}_i\}_{i=1}^{C}, \qquad \hat{\mathbf{F}}_{cir} = \hat{\mathcal{V}}_{cir} \otimes \mathbf{F} = \{\hat{v}_{cir,i} \cdot \mathbf{F}_i\}_{i=1}^{C}. \tag{6.2}$$

To guide feature separation on convolutional layer, we enforce $\hat{\mathbf{F}}_{cr}$ and $\hat{\mathbf{F}}_{cir}$ to pass through a series of operations including **Pooling**, **FC**, **ReLU** and **FC** to synthesize high-level semantic features h_{cr} and $h_{cir} \in \mathbf{R}^{d\times 1}$, where d is the dimension of feature. The learned representation h_{cr} as domain-invariant feature should be fed into the corresponding classifier to promote its discriminative ability. Considering that h_{cir} is required to preserve classification-irrelevant information, the concatenation of h_{cr} and h_{cir} will be taken as input for decoder including several deconvolutional layers [160] to achieve the reconstruction about the input data, i.e., $\hat{\mathbf{X}} = \mathcal{G}(h_{cr}, h_{cir})$, where $\mathcal{G}$ denotes neural network of decoder. Therefore, the objective function of adaptive feature separation module is written as:

$$\begin{aligned} \min_{\Theta} \quad & \mathcal{L}_c(\mathbf{C}(h_{cr}), Y) + \|\mathbf{X} - \mathcal{G}(h_{cr}, h_{cir})\|_F^2 \\ & h_{cr} \in \{h_{cr}^{r,s}, h_{cr}^{ir,s}, h_{cr}^{ir,t}\}, \quad Y \in \{Y^{r,s}, Y^{ir,s}, Y^{ir,t}\} \\ & h_{cir} \in \{h_{cir}^{r,s}, h_{cir}^{ir,s}, h_{cir}^{ir,t}\}, \quad \mathbf{X} \in \{\mathbf{X}^{r,s}, \mathbf{X}^{ir,s}, \mathbf{X}^{ir,t}\}, \end{aligned} \tag{6.3}$$

where Θ refers to all parameters of model, $\mathbf{C} = \{\mathbf{C}^r, \mathbf{C}^{ir}\}$ represents classifier ($h_{cr}^{ir,s}$ and $h_{cr}^{ir,t}$ share classifier C^{ir}, while classifier C^r is target for $h_{cr}^{r,s}$), $\mathcal{L}_c(\cdot)$ means cross-entropy loss and $\mathcal{G}$ consists of two types: $\mathcal{G}_s$ shared by source domain and $\mathcal{G}_t$ used by target domain. Note that the application of objective function requires the consistence of superscript.

6.2.3 Hybrid Generation

The benefit of adaptive feature separation is to extract more discriminative domain-invariant feature with the guidance of label information. To further eliminate domain shift, we propose hybrid reconstruction strategy capturing the connection across various datasets. In other words, we explore the feature alignment between $\mathcal{D}^{ir,s}$ and $\mathcal{D}^{ir,t}$ as well as the consistence of modality over $\mathcal{D}^{r,s}$ and $\mathcal{D}^{ir,s}$ to reduce cross-domain discrepancy of $\mathcal{D}^{r,s}$ and unavailable $\mathcal{D}^{r,t}$.

According to Section 6.2.2, any input signals passing through corresponding encoder and adaptive feature separation module will be transformed into classification-relevant features and classification-irrelevant ones. Due to the paired relationship between $\mathbf{X}^{ir,s}$ and $\mathbf{X}^{ir,t}$, it is reasonable to assume that there exists high similarity between $h_{cr}^{ir,s}$ and $h_{cr}^{ir,t}$ (i.e. $h_{cr}^{ir,s} \equiv h_{cr}^{ir,t}$) derived from corresponding input data. In terms of such equivalent property, we can assert the decoder $\mathcal{G}_t$ performed on ($h_{cr}^{ir,s}$, $h_{cir}^{ir,t}$) and ($h_{cr}^{ir,t}$, $h_{cir}^{ir,t}$) tend to generate the same result, which is formulated as:

$$\mathcal{G}_t(h_{cr}^{ir,s}, h_{cir}^{ir,t}) \equiv \mathbf{X}^{ir,t} \equiv \mathcal{G}_t(h_{cr}^{ir,t}, h_{cir}^{ir,t}). \tag{6.4}$$

With respect to the decoder of source domain $\mathcal{G}_s$, we can similarly draw the conclusion as:

$$\mathcal{G}_s(h_{cr}^{ir,s}, h_{cir}^{ir,s}) \equiv \mathbf{X}^{ir,s} \equiv \mathcal{G}_s(h_{cr}^{ir,t}, h_{cir}^{ir,s}). \tag{6.5}$$

To this end, the loss function of hybrid reconstruction and feature alignment is defined:

$$\begin{aligned} \mathcal{L}_{hr}^{ir} = \lambda \| h_{cr}^{ir,s} - h_{cr}^{ir,t} \|_F^2 + \| \mathcal{G}_t(h_{cr}^{ir,s}, h_{cir}^{ir,t}) - \mathbf{X}^{ir,t} \|_F^2 \\ + \| \mathcal{G}_s(h_{cr}^{ir,s}, h_{cir}^{ir,s}) - \mathbf{X}^{ir,s} \|_F^2, \end{aligned} \tag{6.6}$$

where λ is the hyper-parameter controlling the reconstruction and feature alignment. The first term in Eq. (6.6) not only achieves distribution alignment over task-irrelevant datasets, but also gradually eliminates the difference of models on feature learning. Under such condition, even though target dataset $\mathcal{D}^{r,t}$ is unavailable for training stage, the similarity of model effectively facilitates the consistency of feature representation across $\mathcal{D}^{r,s}$ and

$\mathcal{D}^{r,t}$. Meanwhile, hybrid reconstruction loss plays an essential role in achieving the goal of feature separation, which aims to preserve abundant meaningful and discriminative feature in classification-relevant representation via the last two terms.

From Figure 6.1, we observe that classification-irrelevant units derived from $\mathbf{X}^{r,s}$ and $\mathbf{X}^{ir,s}$ ideally should maintain high correlation, since their corresponding input signals belong to the same modality. However, $h_{cr}^{r,s}$ and $h_{cr}^{ir,s}$ tend to describe distinctive objects of images. The expected association between $\mathcal{D}^{r,s}$ and $\mathcal{D}^{ir,s}$ is expressed as:

$$\mathcal{G}_s(h_{cr}^{r,s}, h_{cir}^{r,s}) \equiv \mathbf{X}^{r,s} \approx \mathcal{G}_s(h_{cr}^{r,s}, h_{cir}^{ir,s}). \tag{6.7}$$

$$\mathcal{G}_s(h_{cr}^{ir,s}, h_{cir}^{ir,s}) \equiv \mathbf{X}^{ir,s} \approx \mathcal{G}_s(h_{cr}^{ir,s}, h_{cir}^{r,s}). \tag{6.8}$$

Therefore, we explore hybrid generation to satisfy such a requirement and reformulate our objective function as:

$$\mathcal{L}_{hr}^{s} = \|\mathcal{G}_s(h_{cr}^{r,s}, h_{cir}^{ir,s}) - \mathbf{X}^{r,s}\|_F^2 + \|\mathcal{G}_s(h_{cr}^{ir,s}, h_{cir}^{r,s}) - \mathbf{X}^{ir,s}\|_F^2. \tag{6.9}$$

Remarks: If we have access to the missing target modality $\mathbf{X}^{r,t}$, the constraint of Eq. (6.9) enables the model to capture relationships: $\mathcal{G}_t(h_{cr}^{r,t}, h_{cir}^{ir,t}) \approx \mathbf{X}^{r,t} \equiv \mathcal{G}_t(h_{cr}^{r,t}, h_{cir}^{r,t})$ and $\mathcal{G}_t(h_{cr}^{ir,t}, h_{cir}^{r,t}) \approx \mathbf{X}^{ir,t} \equiv \mathcal{G}_t(h_{cr}^{ir,t}, h_{cir}^{ir,t})$. Moreover, under the supervision of Eq. (6.9), we also achieve the conclusion $\mathcal{G}_t(h_{cr}^{r,s}, h_{cir}^{r,t}) \approx \mathbf{X}^{r,t} \equiv \mathcal{G}_t(h_{cr}^{r,t}, h_{cir}^{r,t})$ and $\mathcal{G}_s(h_{cr}^{r,t}, h_{cir}^{r,s}) \approx \mathbf{X}^{r,s} \equiv \mathcal{G}_s(h_{cr}^{r,s}, h_{cir}^{r,s})$. Through such mediate manner, the model finally achieves domain adaptation across $\mathcal{D}^{r,s}$ and $\mathcal{D}^{r,t}$.

6.2.4 Training and Inference

Given accessible datatsets $\mathcal{D}^{r,s}$, $\mathcal{D}^{ir,s}$ and $\mathcal{D}^{ir,st}$, we firstly perform initial feature separation within each dataset. And then hybrid reconstruction as an important component captures delicate association across all datasets to gradually reduce cross-domain discrepancy between $\mathcal{D}^{r,s}$ and missing target dataset $\mathcal{D}^{r,t}$. Finally, we utilize the feature extractor of target domain to learn feature of $\mathbf{X}^{r,t}$ and apply classifier $\mathbf{C}_s(\cdot)$ to perform classification task. Therefore, the overall process is summarized as three steps:

Step A: Input data including $\mathbf{X}^{r,s}$, $\mathbf{X}^{ir,s}$ and $\mathbf{X}^{ir,t}$ first is fed into the encoder to learn convolutional features. And then we perform adaptive feature separation on convolutional layers to obtain classification-relevant unit h_{cr} and h_{cir}. Finally, the concatenation of h_{cr} and h_{cir} is exploited to reconstruct input signal. During learning stage, we explore objective function (6.3) to update model.

Step B: To achieve the expected feature separation and domain adaptation, we should integrate hybrid reconstruction and the guidance of label information into a unified loss function as:

$$\begin{aligned} \min_{\Theta} \quad & \mathcal{L}_{hr}^{ir} + \mathcal{L}_{hr}^{s} + \mathcal{L}_c\big(\mathbf{C}(h_{cr}), Y\big) \\ & h_{cr} \in \{h_{cr}^{r,s}, h_{cr}^{ir,s}, h_{cr}^{ir,t}\}, Y \in \{Y^{r,s}, Y^{ir,s}, Y^{ir,t}\}, \end{aligned} \tag{6.10}$$

where $\mathbf{C}$ consists of $\mathbf{C}_s$ classifier used by $h_{cr}^{r,s}$ and $\mathbf{C}_t$ classifier shared by $h_{cr}^{ir,s}$ and $h_{cr}^{ir,t}$. We train the network according to Eq. (6.10) until convergence.

Step C: During inference stage, instances $X^{r,t}$ will be passed through the encoder used by $X^{ir,t}$ to obtain high-level feature $h_{cr}^{r,t}$. Eventually, we utilize classifier $\mathbf{C}_s$ to predict the annotation of $h_{cr}^{r,t}$.

6.3 Experiments

6.3.1 Datasets and Comparisons

We perform experiments on three popular benchmarks involving MNIST [161], Fashion MNIST [162] and EMNIST [163] to verify the effectiveness of our method. For the convenience and clarity, we utilize dataset IDs D_M, D_F and D_E to refer to them. In addition, there exists three techniques to transform each gray-scale image into the corresponding negative, color and edge images.

MNIST (D_M) dataset is developed to identify handwritten digit image. The dataset includes 70,000 gray-scale images, where 60,000 training instances and 10,000 testing images. Each visual instance with same size 28 × 28 only represents one of ten digits from 0 to 9.

Fashion MNIST (D_F) dataset includes abundant fashion trappings images. Experts in fashion field artificially divide them into ten categories: *T-shirt*, *trouser*, *pullover*, *dress*, *coat*, *sandals*, *shirt*, *sneaker*, *bag*, and *ankle boot*. The dataset has the same sample scale with MNIST, i.e 60,000 training instances and 10,000 testing samples. The image size of each sample is also 28×28.

EMNIST (D_E) dataset different from MNIST records extensive handwritten alphabets images. The uppercase and lowercase letters are merged into a balanced dataset with 26 categories. The image size of each sample is 28×28. Moreover, it involves 124,800 images for training and 20,800 images for testing.

Modality Transformation: All instances in the above mentioned datasets are gray-scale images and we define this modality as *G-domain*. To perform domain adaptation, We firstly follow the operations in [159] to convert all original data into negative image (*N-domain*) by using $\mathbf{X}_n = 255 - \mathbf{X}$, $\mathbf{X} \in \mathbf{R}^{m \times n \times 1}$ where m and n are the spatial dimensions of image. Moreover, we apply canny detector to create edge images $\mathbf{X}_e$ (*E-domain*). Finally, in terms of color version, we randomly extract several patches ($\mathbf{P} \in \mathbf{R}^{m \times n}$) from the BSDS500 dataset [164] and then blend them with images $\mathbf{X}$ to form color images $\mathbf{X_c}$ (*C-domain*).

Comparisons: To evaluate the performance of our method, we select three baselines as competed methods which are currently the only works exploring the application of deep learning on zero-shot domain adaptation problem. The first compared approach is *ZDDA* [156], which propose sensor fusion to solve domain shift. Moreover, [159] utilizes two models named $CoGAN$ and $CoCoGAN$ to address ZSDA issue, which are considered as two various approaches.

6.3.2 Implementation Details

The network architecture of our method mainly includes three components: encoder, decoder and classifier. Although source and target utilize various networks, they have the same network structure. Thus, we take the branch of source domain as an example to illustrate the specific implementation. With respect to the encoder, we adopt three convolutional layers with stride 2 to extract channel-level feature and apply **ReLU** to activate the output of the first two layers. Symmetrically, the decoder has three deconvolutional layers with stride 2 to recover hidden representation to input data. There are two classifiers used in our proposed method and they both have two full-connection layers followed by

Table 6.1: Classification Accuracy (%) of our method and three baselines for domain adaptation from gray-scale modality (*G-domain*) to color modality (*C-domain*). The best result in each column is in bold.

RT	MNIST (D_M)		Fashion-Mnist		EMNIST (D_E)	
IRT	D_F	D_E	D_M	D_E	D_M	D_F
ZDDA [156]	73.2	94.8	51.6	65.3	71.2	47.0
CoGAN [159]	68.3	74.7	39.7	55.8	46.7	41.8
CoCoGAN [159]	78.1	**95.6**	56.8	66.8	**75.0**	54.8
HGNet	**85.3**	95.0	**64.5**	**71.1**	71.3	**57.9**

Softmax function.

6.3.3 Experimental Results

In order to validate the effectiveness of our method, we create five different zero-shot domain adaption settings. We firstly consider gray-scale images as source domain and the other three domains will be target domain. Thus, there are three domain adaptation tasks: *G-domain* → *N-domain*, *G-domain* → *E-domain* and *G-domain* → *C-domain*. In addition, we also attempt to transfer knowledge from color domain or negative domain to gray domain, i.e., *C-domain* → *G-domain* and *N-domain* → *G-domain*.

According to descriptions of dataset, we know these three datasets involves three completely distinctive objects: digits, trappings and letters. When selecting one of them as task-relevant dataset, we can consider others as task-irrelevant datasets which assist model to capture cross-domain discrepancy and promote classification accuracy on missing target modality ($D^{r,t}$). Firstly, we attempt to transfer knowledge from gray-scale modality (*G-domain*) to color modality (*C-color*). Compared with gray-scale image, original RGB image generally involve three color channels, which dramatically increase the difficult in achieving domain adaptation. Experimental performances are summarized in Table 6.1. In terms of these results, our proposed method (HGNet) obtains the best classification accuracy in three datasets. And there exist significant differences between HGNet and CoCoGAN achieving the second best performance. Specifically, our proposed approach surpasses CoCoGAN by

Table 6.2: Classification Accuracy (%) of our method and three baselines for two domain adaptation tasks :*N-domain* → *G-domain* and *G-domain* → *N-domain*. The best result in each column is in bold.

Task	*N-domain*→*G-domain*						*G-domain*→*N-domain*			
RT	D_M		D_F		D_E		D_M		D_F	D_E
IRT	D_F	D_E	D_M	D_E	D_M	D_F	D_F	D_E	D_E	D_F
ZDDA [156]	78.5	87.6	56.6	67.1	67.7	45.5	77.9	90.5	62.7	53.4
CoGAN [159]	66.1	76.3	49.9	58.7	53.0	32.5	62.7	72.8	51.2	39.1
CoCoGAN [159]	80.1	93.6	63.4	72.8	**78.8**	58.4	80.3	93.1	69.3	56.5
HGNet	**87.5**	**95.0**	**64.6**	**75.1**	78.0	**67.9**	**83.7**	**95.7**	**71.7**	**62.3**

7.7% when Fashion-MNIST and MNIST separately are task-relevant and task-irrelevant datasets. On the one hand, the empirical results provide convincing answer (No) to the question in Section 3.1: is the generation of missing target dataset necessary for learning domain-invariant feature. On the other hand, it illustrates that hybrid generative manner guarantees the uniqueness of feature separation and the application of it enable model to learn more discriminative domain-invariant feature.

For the second step, we conduct transformation between gray-scale modality (*G-domain*) and negative modality (*N-domain*) and summary the corresponding performances in Table 6.2. From these experimental results, we can obtain three conclusions. First of all, the proposed algorithm (HGNet) achieves more promising performances than other baselines in most cases. Specifically, when separately selecting D_E and D_F as task-relevant and task-irrelevant datasets, our approach outperforms CoCoGAN by 5.8% on the domain adaptation task (*G-domain* → *N-domain*). Secondly, we notice that classification accuracy of all mentioned methods on Fashion-MNIST (task-relevant datatset) is lower than that on other two datasets. The main reason for this derives from that trappings images are more complex than digits and letters images. However, HGNet still improve 1%∼3% when compared with the second best result obtained by CoCoGAN. Finally, although these two transformation (*G-domain* → *N-domain* and *N-domain* → *G-domain*) are mutually inverse operations, classification accuracy of most approaches on *G-domain* → *N-domain* are better

Table 6.3: Classification Accuracy (%) of our method and three baselines for two domain adaptation tasks :*G-domain* → *E-domain* and *C-domain* → *G-domain*. The best result in each column is in bold.

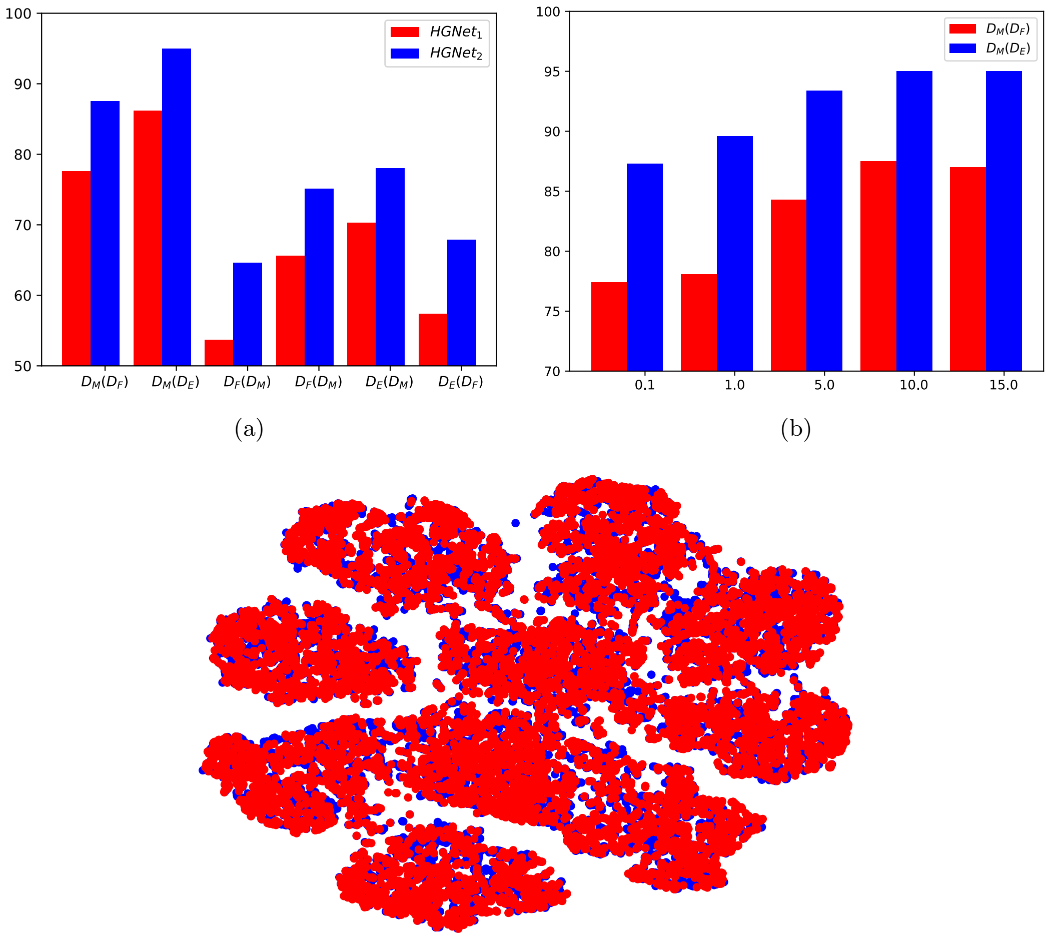
Task	*G-domain*→*E-domain*				*C-domain*→*G-domain*			
RT	MNIST (D_M)		EMNIST (D_E)		MNIST (D_M)		Fashion (D_F)	
IRT	D_F	D_E	D_M	D_F	D_F	D_E	D_M	D_E
ZDDA [156]	72.5	93.2	73.6	50.7	67.4	87.6	55.1	59.5
CoGAN [159]	67.1	81.5	63.6	51.9	54.7	63.5	43.4	51.6
CoCoGAN [159]	79.6	95.4	77.9	58.6	73.2	94.7	61.1	**70.2**
HGNet	**86.5**	**96.1**	**81.1**	**59.5**	**78.9**	**95.0**	**65.9**	68.5

than their performances on *N-domain* → *G-domain*. But the results of HGNet on these two transformations are competitive, which means our method has much better generalization.

In the final experiment, we explore *G-domain*→*E-domain* and *C-domain*→*G-domain* to further verify the effectiveness of HGNet. Results are reported in Table 6.3. The performance of HGNet is better than others in most cases. Interestingly, we find that although there exists high similarity between D_M and D_E, it difficult for most methods to achieve great transformation on D_E with the assistance of D_M. Different from them, our method fully utilizes association across all available datasets to reduce cross-domain discrepancy, leading to the improvement on classification accuracy to 81.1%.

6.3.4 Ablation Study

Effect of Hybrid Strategy: According to the discussion about hybrid reconstruction, we know that this part enable the proposed model to further guarantee the uniqueness of feature separation and promote generalization across various domains by using association of all given datasets. In order to clearly observe the effect of hybrid reconstruction, we firstly attempt to remove this part from our method to form another competed method named as HGNet$_1$, while the overall version of our method is denoted as HGNet$_2$. The goal of experiments in this section is to achieve the transformation from *N-domain* to *G-domain* and Figure 6.2 (a) lists results, where the expression *A(B)* means *A* is task-relevant dataset



(a)

(b)

(c)

Figure 6.2: Experiments are performed on adaptation from *N-domain* to *G-domain*. And the expression *A(B)* means A is the task-relevant dataset while B represents the task-irrelevant one. (a) We denote our proposed method without hybrid reconstruction as $HGNet_1$ and the overall version as $HGNet_2$. (b) We select λ from $\{0.1, 1.0, 5.0, 10.0, 15.0\}$ and observe the classification accuracy. (c) When D_E is the task-irrelevant datasets, we show the feature visualization on MNIST.

while B represents task-irrelevant one.

As seen in Figure 6.2 (a), the absence of hybrid reconstruction suffers from significant negative influence on the classification accuracy. HGNet$_2$ outperforms HGNet$_1$ by 10% for $D_F(D_M)$, illustrating that hybrid strategy not only effectively generates more discriminative feature representation but also captures more cross-domain information from all available data to reduce domain shift.

Additionally, we present the generated images in Figure 7.4 via hybrid generation to verify its ability performing transformation between source and target domains. In terms of

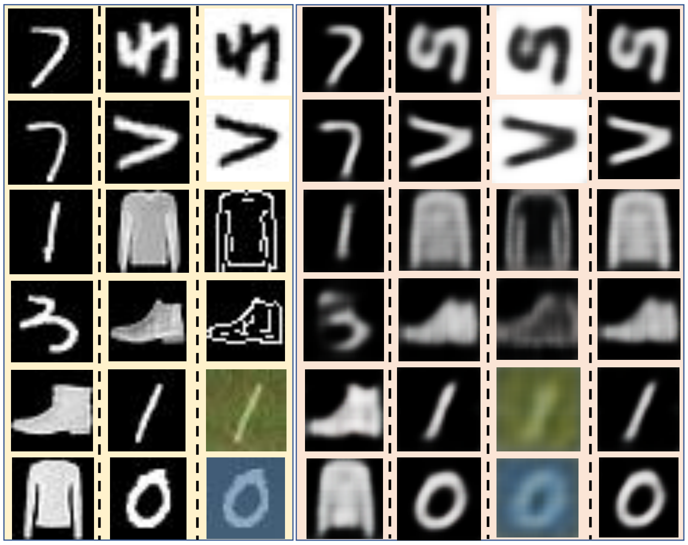
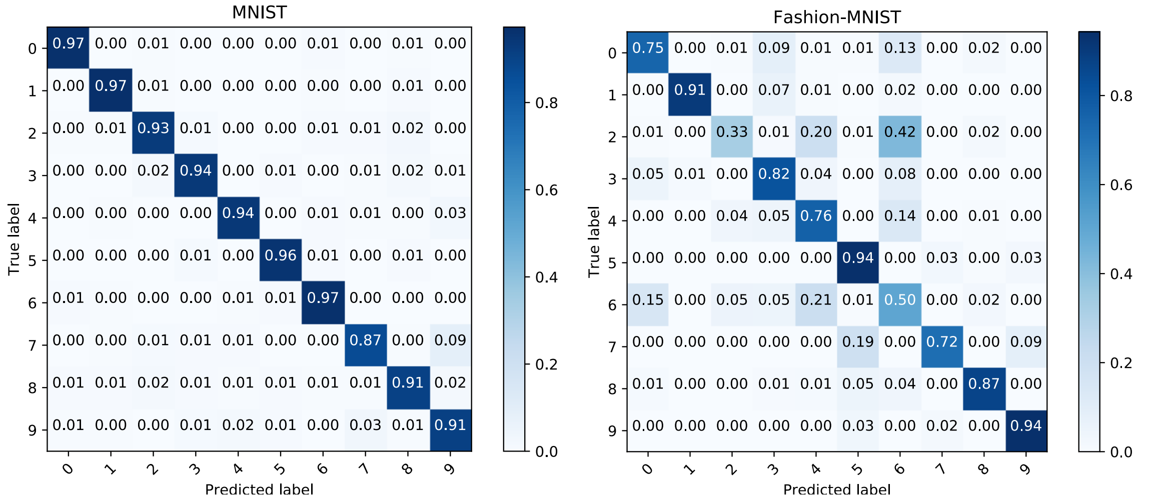


Figure 6.3: Visualization of hybrid generation. The first three columns represents the inputs: $X^{r,s}$, $X^{ir,s}$ and $X^{ir,t}$, while the last four columns are hybrid generative visual signals: $\mathcal{G}_s(h_{cr}^{r,s}, h_{cir}^{ir,s})$, $\mathcal{G}_s(h_{cr}^{ir,s}, h_{cir}^{r,s})$, $\mathcal{G}_t(h_{cr}^{ir,s}, h_{cir}^{ir,t})$ and $\mathcal{G}_s(h_{cr}^{ir,t}, h_{cir}^{ir,s})$.

the visualization, we find that hybrid strategy captures cross-domain discrepancy. Specifically, in the first two rows, images synthesised by $\mathcal{G}(h_{cr}^{ir,s}, h_{cir}^{ir,t})$ actually integrate main objects from $X^{ir,s}$ and the corresponding modality style (*N-domain*) from $X^{ir,t}$. It means that our proposed method achieves high-quality separation of semantic information, which assists model to learn domain-invariant feature and promote classification accuracy.

Parameters Analysis: To show the function of feature alignment on task-irrelevant dataset, we change the value of λ from *0.1* to *15* and record results (*N-domain*→*G-domain*) in Figure 6.2 (b). With the increasing of λ, HGNet achieves higher accuracy, illustrating that such feature alignment manner has positive effect on solving the domain shift issue on task-relevant dataset.

Visualization of Latent Space: To further analyse distribution of high-level feature, we draw feature visualization and confusion matrix on MNIST and Fashion-MNIST in Figure 6.2 (c) and Figure 6.4. For these experiments, we select EMNIST as task-irrelevant datasets and transfer negative images (*N-domain*) into gray-scale modality (*G-domain*). From the performance, we know that HGNet learns clear boundary between various categories, which significantly promotes feature discriminative.



MNIST

True label \ Predicted label	0	1	2	3	4	5	6	7	8	9
0	0.97	0.00	0.01	0.00	0.00	0.00	0.01	0.00	0.01	0.00
1	0.00	0.97	0.01	0.00	0.00	0.00	0.00	0.00	0.01	0.00
2	0.00	0.01	0.93	0.01	0.00	0.00	0.01	0.01	0.02	0.00
3	0.00	0.00	0.02	0.94	0.00	0.01	0.00	0.01	0.02	0.01
4	0.00	0.00	0.00	0.00	0.94	0.00	0.01	0.01	0.00	0.03
5	0.00	0.00	0.00	0.01	0.00	0.96	0.01	0.00	0.01	0.00
6	0.01	0.00	0.00	0.00	0.01	0.01	0.97	0.00	0.00	0.00
7	0.00	0.00	0.01	0.01	0.01	0.00	0.00	0.87	0.00	0.09
8	0.01	0.01	0.02	0.01	0.01	0.01	0.01	0.01	0.91	0.02
9	0.01	0.00	0.00	0.01	0.02	0.01	0.00	0.03	0.01	0.91

(a) Confusion Matrix on MNIST

Fashion-MNIST

True label \ Predicted label	0	1	2	3	4	5	6	7	8	9
0	0.75	0.00	0.01	0.09	0.01	0.01	0.13	0.00	0.02	0.00
1	0.00	0.91	0.00	0.07	0.01	0.00	0.02	0.00	0.00	0.00
2	0.01	0.00	0.33	0.01	0.20	0.01	0.42	0.00	0.02	0.00
3	0.05	0.01	0.00	0.82	0.04	0.00	0.08	0.00	0.00	0.00
4	0.00	0.00	0.04	0.05	0.76	0.00	0.14	0.00	0.01	0.00
5	0.00	0.00	0.00	0.00	0.00	0.94	0.00	0.03	0.00	0.03
6	0.15	0.00	0.05	0.05	0.21	0.01	0.50	0.00	0.02	0.00
7	0.00	0.00	0.00	0.00	0.00	0.19	0.00	0.72	0.00	0.09
8	0.01	0.00	0.00	0.01	0.01	0.05	0.04	0.00	0.87	0.00
9	0.00	0.00	0.00	0.00	0.00	0.03	0.00	0.02	0.00	0.94

(b) Confusion Matrix on Fashion-MNIST

Figure 6.4: Visualization of Confusion Matrix. Experiments are performed on adaptation from *N-domain* to *G-domain*. For these two experiments, we select EMNIST as the task-irrelevant datasets.

6.4 Conclusion

Zero-shot Domain Adaptation (ZSDA) assumes that we hardly access target samples during training stage. To fight off ZSDA more effectively, we propose a novel approach named Hybrid Generative Network (HGNet) including feature extractor, adaptive feature separation module, hybrid generator and classifier. Concretely, feature extractor learns representations from visual signals, and then adaptive feature separation module distinguishes classification-relevant units from classification-irrelevant ones storing meaningless semantic information. Moreover, we adopt two manners to perform high-quality feature separation. One is to use annotation as supervision to generate discriminative feature. Another is to exploit hybrid generative strategy to extract association across various available datasets. Finally, extensive experimental results validate the effectiveness of HGNet on solving ZSDA problem.

Chapter 7

Representation Fusion for Incomplete Multi-View Domain Adaptation

7.1 Background

Deep neural network (DNN) recently becomes the dominate technique in computer vision community due to its success on the real-world applications such as image classification [165, 165, 166], object detection [167] and image segmentation [168, 169]. As a data-driven learning strategy, DNN generally requires considerable training samples with high-quality annotations to capture the intrinsic semantic knowledge. However, the data collection and manual annotation tend to be expensive and time-consuming [11, 170, 171]. To benefit from external resources, recent solutions pay more attentions to transfer learning, especially for unsupervised domain adaptation (UDA) [30, 63, 94, 172].

UDA aims to transfer well-supervised source knowledge to assist the specific tasks in target domain without any annotation information [96, 173]. However, data collection typically occurring in varying environments easily triggers the significant distribution discrepancy across source and target samples [15, 174]. The main challenge of UDA is how to learn domain-invariant feature representations. Along with this direction, the UDA algorithms mainly explore metric-based scheme and adversarial training fashion. Specifically, one of the classical and effective metric-based strategies transforms target samples into source latent space and explore their sample-wise association to eliminate domain mismatch [175]. However, the alignment method needs to observe all data to accurately estimate the relation of source and target instances, which difficultly adjusts to the mini-batch training manner in DNN. In addition, the basic UDA setting considers that the images of source and target domain are merely captured by one sensor. But the practical application always deploys multiple sensors such as the autonomous vehicle to obtain more sufficient information to boost the model performance.

A few efforts [25,26] have explored multi-view domain adaptation (MVDA), where source and target data are both collected from multiple sensors. The intuitive idea is to convert MVDA into a UDA problem by independently aligning source and target instances within each view and fusing multi-view semantic information within individual domain. They have achieved promising performance on solving MVDA and abundant empirical studies illustrate that the simple alignment-and-fusion promotes model performance on identifying target samples with more enriched data collected by multiple sensors. However, equipment rehabilitation to upgrade previous single-sensor devices with multiple sensors causes additional cost overhead, which makes MVDA to be invalid for several practical application scenarios. Instead, we post a question that *"Can we develop more effective domain adaptation algorithms to benefit single-sensor target data from enriched source data with multiple sensors?"*. This problem is defined as incomplete multi-view domain adaptation (**IMVDA**), where there are multi-view complete data in source domain, while single-view instances in target domain. This problem is under insufficient exploration in the literature.

To overcome IMVDA challenge, we propose a novel method named Channel Enhancement and Knowledge Transfer (**CEKT**) shown in Figure 7.1, which not only conducts multi-view semantic fusion within source domain but also transfers the integrated knowledge for the use of target domain. Concretely, CEKT explores the sparse attribution of channel to distinguish view-common from view-specific feature maps and exchanges view-specific channels across multiple views to fuse their semantic information. Furthermore, we develop a metric of channel similarity to highlight the representation of significant channels,

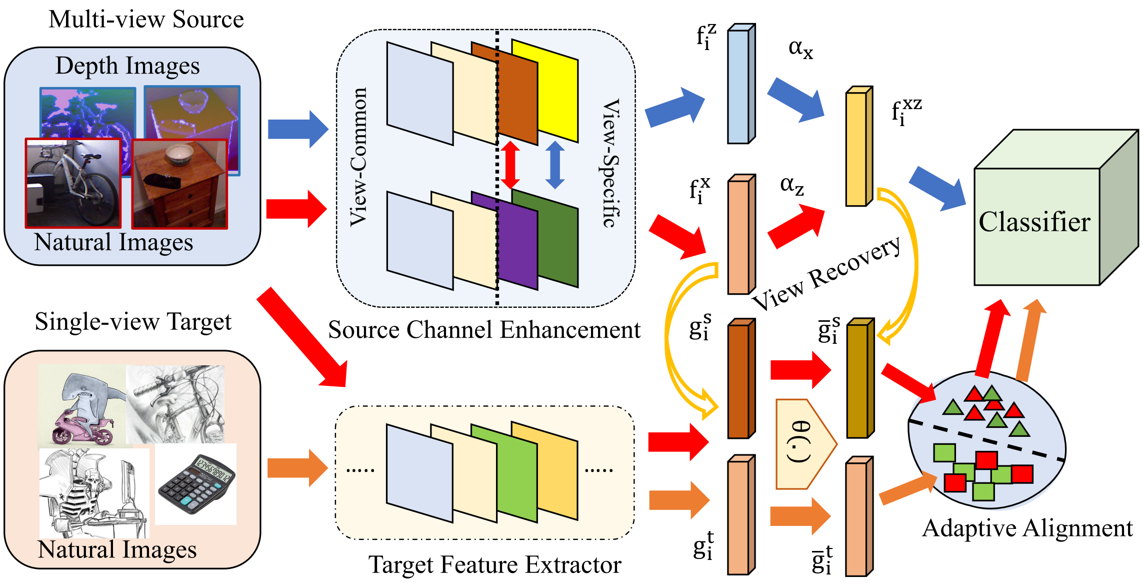


Figure 7.1: Overview of our channel enhancement and knowledge transfer framework (CEKT) for incomplete multi-view domain adaptation (IMVDA). Specifically, the source channel enhancement module distinguishes view-common from view-specific channels and explores the channel similarity to emphasis essential representation. The source triggered missing view recovery teaches target model how to generate multi-view knowledge. And the adaptive alignment module aims to eliminate domain mismatch within the identical subspace.

which assists model learning with more discriminative features. Moreover, we introduce a parallel target model taking source and target samples from the same view as input. The source model trained in the first step teaches the target model to produce multi-view semantic only with single view data. In addition, we propose a novel adaptive subspace alignment to gradually mitigate domain discrepancy in an end-to-end training manner. To sum up, the main contributions of this work are highlighted in three folds:

- First, our proposed CEKT introduces a novel channel enhancement mechanism to preserve considerable view-common semantic knowledge and exchange view-specific semantic to enrich the representation of each view. This module not only effectively achieves feature fusion but also emphasizes more discriminative features for the classification task.

- Second, the adaptive knowledge transfer module explores the supervision of source model to supervise the target model to approximate multi-view semantic information, which mitigates the negative influence of missing view on target domain. Simultaneously, we present a novel adaptive subspace alignment method to learn domain-invariant representations.

- Finally, we exploit many public-available real-world image datasets to imitate the IMVDA scenario and conduct abundant experiments to evaluate the performance of our CEKT. The corresponding experimental results and analysis fully demonstrate the effectiveness of our method.

7.2 Proposed Method

7.2.1 Preliminary & Motivation

Formally for the IMVDA problem, we are given the well-annotated source domain with enriched views[1] as $\mathcal{D}_s = \{(\mathbf{x}_i^s, \mathbf{z}_i^s, y_i)\}_{i=1}^{n_s}$ and the unlabeled target domain with only single view as $\mathcal{D}_t = \{\mathbf{x}_i^t\}_{i=1}^{n_t}$, where $\mathbf{x}$ and $\mathbf{z}$ denote two view-paired samples, y represents the corresponding source label, and n_s and n_t are the number of source and target samples, respectively. The goal of IMVDA is to transfer the enriched view information and well-annotated label information in the source domain to improve the single-view target recognition.

Therefore, two-fold challenges should be considered: 1) How to effectively integrate multi-view semantics to boost performance of model, and 2) How to transfer knowledge from multi-view source domain to single view target one. To address these questions, we propose a novel solution named Channel Enhancement and Knowledge Transfer (CEKT) framework as Figure 7.1. Concretely, CEKT involves two components, i.e., a source channel enhanced network and an adaptive knowledge transfer network. The former one aims to distinguish view-common channels from view-specific channels where semantic fusion occurs and exploit cross-view channel similarity to enhance the representation of necessary channels. The latter one attempts to adaptively learn a target-to-source projection to mitigate the domain mismatch.

[1] This paper considers the case that the source domain contains two views while target domain includes only single view.

7.2.2 Source Channel Enhanced Network

Cross-view Channel Enhancement. Batch normalization (BN) [176] is widely used in deep neural networks to scale the hidden features of the specific layer to accelerate convergence and avoid the model collapse as:

$$\hat{h}_c = \gamma_c \frac{(h_c - \mu_c)}{\sqrt{\sigma_c^2 + \epsilon}} + \beta_c, \tag{7.1}$$

where h_c, $\hat{h}_c$ mean the input and output of the BN module, μ_c, σ_c are the mean and variance of the c-th channel, and γ_c, β_c are trainable parameters. However, from the perspective of channel exchange [177], the model training gradually neglects the representation of task-irrelevant channels as $\gamma_c \to 0$, and multi-view data cause the channels $(h_{x,c},\ h_{z,c})$ from $(\mathbf{x}^s, \mathbf{z}^s)$ to be activated differently. Then, Wang et. al. proposed channel exchange for two views to compensate each other as [177]:

$$\hat{h}_{x/z,c} = \gamma_{z/x,c} \frac{(h_{z/x,c} - \mu_{z/x,c})}{\sqrt{\sigma_{z/x,c}^2 + \epsilon}} + \beta_{z/x,c}, \quad \text{if} \quad \gamma_{x/z,c} < \delta, \tag{7.2}$$

where δ is an adjustable threshold, and a sparse regularization term $\sum_{c=1}^{C} |\gamma_{x/z,c}|$ is introduced to encourage more channel exchanges. Such channel exchange totally relies on the learned $\gamma_{x/z,c}$, which makes channel exchange in an unsupervised fashion without considering sharing channels across views.

Thus, we develop a Cross-view Channel Enhancement (C^{2}E) module. Specifically, for one concrete layer, all channels are divided into two groups: view-common channels and view-specific ones. Under this condition, we suppose view-common channels tend to involve considerable shared semantics, where the corresponding parameters $\gamma_{x/z,c}$ should be compact rather than sparse, and view-specific channels carry the unique information for each view and should be exchanged and enhanced. With this consideration, the ℓ_1-norm over the parameters is a promising manner to highlight the difference across view-specific channels. In implementation, we consider the first half of all feature maps as the view-common channels and the remaining ones as view-specific parts. Thus, we adopt the following constraint for parameters $\gamma_{x/z}$ as:

$$\min_{\gamma_{l,c}} \mathcal{L}_\gamma = \sum_{l=1}^{L} \left(\sum_{c=1}^{\lfloor C/2 \rfloor} \gamma_{l,c}^2 + \sum_{c=\lfloor C/2 \rfloor}^{C} |\gamma_{l,c}| \right), \tag{7.3}$$

where we omit the superscript (x, z) for convenience, C and $\lfloor C/2 \rfloor$ mean the number of channel and the rounding or flooring operation, and L is the number of network layers attached with the BN module. It is worth noting that only the view-specific channels participate in the channel exchange via Eq. (7.2). Through the above strategy, we not only achieve feature fusion but also preserve as much view-common semantics as possible. Hence, $\gamma_{x/z,c} \geq \delta$ illustrates that this channel can contribute to the classification task.

To further enhance the channels shared across views, we propose a strategy to identify those channels and amplify their presence during batch normalization. As two views data present the identical content in various forms, their representations to the necessary information such as the contour of object tend to be similar or even consistent. In other words, the c-th channel with a high similarity across two views should be considered as an important component with a high confidence. Thus, the similarity (s_c) of two views at channel c is defined as:

$$s_c = \frac{\exp(-\|\mu_{x,c} - \mu_{z,c}\|_2/\eta)}{\sum\limits_{c=1}^{C} \exp(-\|\mu_{x,c} - \mu_{z,c}\|_2/\eta)}, \tag{7.4}$$

where $\sum_c s_c = 1$ and η controls the change of scale. Then, we first adjust the importance of channel with $\hat{h}_{x/z,c} = (1 + s_c)\hat{h}_{x/z,c}$ before the channel exchange in Eq. (7.2). For instance, when the two channels are very different, corresponding s_c plays a small fraction of the similarity vector and, hence, the importance of the c-th channel is not augmented with a relatively small s_c.

Data-dependant Cross-view Fusion. For now, our module is easily applied into most deep neural network $\mathcal{F}(\cdot)$ mapping the original image into the high-level features $\mathbf{f}_x = \mathcal{F}(\mathbf{x})$ or $\mathbf{f}_z = \mathcal{F}(\mathbf{z})$. To further learn robust features, we adopt a data-dependant fusion manner to obtain these high-level representations as:

$$\mathbf{f}_{xz} = \alpha_x \mathcal{F}(\mathbf{x}) + \alpha_z \mathcal{F}(\mathbf{z}), \tag{7.5}$$

where $\alpha_{x/z}$ are the probability score for two views and we plug in the softmax layer $\mathbf{s}(\cdot)$ to $\mathcal{F}(\mathbf{x})$ and $\mathcal{F}(\mathbf{z})$ to learn the data-dependant fusion weights.

Finally, the multi-class source classifier $\mathcal{C}(\cdot)$ takes the fused features as input to generate the prediction. The objective function for training the source model is formulated as:

$$\min_{\mathcal{F},\mathcal{C},\mathbf{s},\gamma} \mathcal{L}_s = \sum_{i=1}^{n_s} \mathcal{L}_c\Big(\mathcal{C}(\mathbf{f}_{xz}^i), y_i\Big) + \lambda_\gamma \mathcal{L}_\gamma, \tag{7.6}$$

where λ_γ is a trade-off parameter and $\mathcal{L}_c(\cdot,\cdot)$ is the classical cross-entropy loss.

7.2.3 Adaptive Knowledge Transfer Network

The target domain lacks one view and exists considerable distribution difference with source domain, which makes it unreasonable to directly identify target samples with multi-view source model. Thus, the current challenge is how to effectively transfer source fused knowledge to the target domain. Along with this direction, we construct a novel adaptive knowledge transfer network (AKT), whose core is to associate two domains with source view data x_i^s as the bridge. Concretely, we introduce an additional target network $\mathcal{G}(\cdot)$ with the same network architecture to source and the conventional BN module.

Source Triggered Missing View Recovery. To guide the target network with the ability for missing view, we allow source sample $\mathbf{x}_i^s$ and target sample $\mathbf{x}_j^t$ to pass through the target network $\mathcal{G}(\cdot)$ so that we can obtain the high-level features, i.e., $\mathbf{g}_i^s = \mathcal{G}(\mathbf{x}_i^s)$ and $\mathbf{g}_j^t = \mathcal{G}(\mathbf{x}_j^t)$. Following that, we deploy one dimensionality-identical full-connection layer with trainable parameter θ to obtain $\bar{\mathbf{g}}_i^s$ and $\bar{\mathbf{g}}_j^t$, which aims to recover the missing view information for the target network by mapping one view to two-view fused representation.

Since the target model does not directly touch $\mathbf{z}_i^s$, we expect to learn the fused semantic only with one source view data. As DNN manifests strong approximation capability by using the convolution layers and non-linear mapping [178], it fits better to the given target. Inspired by this observation, when accessing the fused representations with fixed source model, we make $\mathbf{g}_i^s$ and $\bar{\mathbf{g}}_i^s$ approximate $\mathbf{f}_x^i$ and $\mathbf{f}_{xz}^i$, respectively, to mimic the fused semantic features. Hence, we propose a source triggered missing view recovery term as:

$$\min_{\mathcal{G},\theta} \mathcal{L}_g = \sum_{i=1}^{n_s} \Big(\|\mathbf{g}_i^s - \mathbf{f}_x^i\|_2^2 + \|\bar{\mathbf{g}}_i^s - \mathbf{f}_{xz}^i\|_2^2 \Big). \tag{7.7}$$

In this way, the source model teaches the target one to offset the absence of the other view. Moreover, as $\mathbf{x}^s$ and $\mathbf{x}^t$ belong to the same view, the imitative manner brings semantics of the other view to feature learning of target samples. Certainly, the significant domain shift across $\mathbf{x}^s$ and $\mathbf{x}^t$ obstructs the delivery of additional semantic knowledge to the target domain. Thus, the target model needs to achieve distribution alignment by gradually eliminating the cross-domain discrepancy.

Adaptive Cross-Domain Alignment. The direct alignment approach is first to transform all source and target instances into the shared latent space and then to reduce the sample-wise distance with the manifold theory. The formulation of this classical strategy [175] is:

$$\min_{\mathbf{A}^{st}} \|\bar{\mathbf{G}}^s - \mathbf{A}^{st}\bar{\mathbf{G}}^t\|_{\mathrm{F}}^2 + \Omega(\mathbf{A}^{st}), \tag{7.8}$$

where $\|\cdot\|_{\mathrm{F}}$ denotes the Frobenius norm, $\bar{\mathbf{G}}^{s/t}$ is the feature matrix of all samples $\bar{\mathbf{g}}_i^{s/t}$, and $\mathbf{A}^{st}$ is defined as the transformation matrix mapping target features into the source feature subspace, and $\Omega(\mathbf{A}^{st})$ denotes a regularization term over $\mathbf{A}^{st}$ such as the ℓ_2-norm or ℓ_1-norm. This strategy achieves promising performance on domain adaptation with shallow feature extractors [179]. However, the feature transformation requires simultaneous access to all samples, which the mini-batch training mechanism used in DNN hardly satisfies. Meanwhile, a direct computation of $\mathbf{A}^{st}$ within each mini-batch is unreasonable since the insufficient samples fail to accurately capture the association of samples. To break the bottleneck, we present an adaptive alignment solution involving two fully connected layers without bias terms. The features $\bar{\mathbf{g}}_i^s$ and $\bar{\mathbf{g}}_j^t$ are fed into it to calculate the similarity $\mathbf{A}_{ij}^{st}$ via:

$$\mathbf{A}_{ij}^{st} = \delta\Big(\langle W_s\bar{\mathbf{g}}_i^s, W_t\bar{\mathbf{g}}_j^t\rangle\Big), \tag{7.9}$$

where $W_{s/t}$ is the projection matrix, $\delta(\cdot)$ denotes an activation function such as ReLU, and $\langle\cdot,\cdot\rangle$ denotes the inner product operation. During the update of $W_{s/t}$, the inputs are fixed.

As the model training, $W_{s/t}$ gradually learns the intrinsic distribution information of overall dataset and can accurately estimate the sample-wise relationship.

On the other hand, we can access to the category probability of sample with $\mathbf{p}_i^{s/t} = \mathcal{C}(\bar{\mathbf{g}}_i^{s/t})$. As $\mathbf{p}_i^{s/t}$ with more discriminative information can reflect the structural relation of hidden features via $\bar{\mathbf{A}}_{ij}^{st} = \langle \mathbf{p}_i^s, \mathbf{p}_j^t \rangle$, we propose the adaptive cross-domain alignment as:

$$\min_{\mathcal{G},\theta,W_{s/t}} \mathcal{L}_a = \|\bar{\mathbf{G}}^s - \mathbf{A}^{st}\bar{\mathbf{G}}^t\|_{\mathrm{F}}^2 + \|\mathbf{A}^{st} - \bar{\mathbf{A}}^{st}\|_{\ell_1}, \tag{7.10}$$

where $\|\cdot\|_{\ell_1}$ denotes the ℓ_1-norm. $\mathbf{A}^{st}$ and $\bar{\mathbf{A}}^{st}$ are normalized along the row dimension. According to the guidance of adaptive similarity $\mathbf{A}^{st}$, the source features can be represented by the similar ones in target domain, and Eq. (7.10) effectively reduces their divergence to mitigate the domain mismatch.

7.2.4 Overall Objective

We first finalize the objective function for the target model. To preserve abundant source knowledge, we adopt source annotations to supervise the target model training. Similar to [44], the pseudo labels of target samples are explored to make target features more discriminative. Specifically, for each epoch, the predictions of target samples (y_j^t) with the fixed target model are used to calculate the class centers, $\mathcal{O}_k = \frac{1}{n_k}\sum_{j=1}^{n_t}\mathbb{I}(y_j^t = k)\bar{\mathbf{g}}_j^t$, where n_k is the number of target samples from the k-th class and $\mathbb{I}(\cdot)$ is the indicator function. With the class centers, the K-means clustering is adopted to reassign the optimized labels $\hat{y}_j^t$ to target samples. The loss function to the target model is defined as:

$$\min_{\mathcal{G},\theta,\mathcal{C},W_{s/t}} \mathcal{L}_t = \mathcal{L}_c^s + \lambda_g\mathcal{L}_g + \lambda_\tau(\mathcal{L}_a + \mathcal{L}_c^t), \tag{7.11}$$

where $\mathcal{L}_c^s$ denotes source supervision loss as $\sum_{i=1}^{n_s}\mathcal{L}_c(\mathcal{C}(\bar{\mathbf{g}}_i^s), y_i^s)$, $\mathcal{L}_c^t$ denotes the pseudo target supervision loss as $\sum_{j=1}^{n_t}\mathcal{L}_c(\mathcal{C}(\bar{\mathbf{g}}_i^t), \hat{y}_j^t)$, and λ_g, λ_τ are trade-off parameters. To avoid the negative effect in the beginning, we define λ_τ as $\frac{1-\exp(-10\tau)}{1+\exp(-10\tau)}$ with the changing of epoch number (τ).

Then, for the overall training strategy, we adopt an iterative training manner to optimize both source and target networks. Concretely, Eq. (7.6) is used to optimize the parameters of source model with the fixed target network $\mathcal{G}(\cdot)$ and then we update target model via Eq. (7.11) with the frozen source network $\mathcal{F}(\cdot)$.

7.2.5 Theoretical Analysis

In Eq. (7.3), we adopt two different constraints on the scaling factors $\gamma_{l,c}$, which enable the network to **actively** learn view-specific and view-common knowledge in various channels, respectively. Similar with [180], we deduce the following theorem to explain why the $\sum_{c=\lfloor C/2 \rfloor}^{C} |\gamma_{l,c}|$ can assist the model to capture view-specific information and the function of $\sum_{c=1}^{\lfloor C/2 \rfloor} \gamma_{l,c}^2$.

Theorem 1. *The proposed $\sum_{c=\lfloor C/2 \rfloor}^{C} |\gamma_{l,c}|$ will definitely make the corresponding scaling factors towards zero with the probability $2\Phi\big(\lambda_\gamma(\frac{\partial \mathcal{L}_c}{\partial \hat{h}_c})^{-1}\big) - 1$, where the $\Phi(\cdot)$ denotes the cumulative probability of standard Gaussian. To be simple, the subscript l of $\gamma_{l,c}$ is mitigated.*

Proof. According to Eq. (7.6), it is straightforward to deduce the derivative of $\mathcal{L}_s$ with respect to $\gamma_c, c \in [C/2, C]$ as the following:

$$\frac{\partial \mathcal{L}_s}{\partial \gamma_c} = \begin{cases} \dfrac{\partial \mathcal{L}_c}{\partial \hat{h}_c} \dfrac{(h_{z/x,c} - \mu_{z/x,c})}{\sqrt{\sigma_{z/x,c}^2 + \epsilon}} + \lambda_\gamma \dfrac{\partial \mathcal{L}_\gamma}{\partial \gamma_c}, & \gamma_c > 0 \\ \dfrac{\partial \mathcal{L}_c}{\partial \hat{h}_c} \dfrac{(h_{z/x,c} - \mu_{z/x,c})}{\sqrt{\sigma_{z/x,c}^2 + \epsilon}} - \lambda_\gamma \dfrac{\partial \mathcal{L}_\gamma}{\partial \gamma_c}, & \gamma_c < 0 \end{cases} \tag{7.12}$$

When the model training approaches convergence, the derivative of $\mathcal{L}_c$ w.r.t $\hat{h}_c$ approximates zero. Due to $\lambda\gamma > 0$, we easily achieve the following inequality:

$$\begin{cases} \dfrac{(h_{z/x,c} - \mu_{z/x,c})}{\sqrt{\sigma_{z/x,c}^2 + \epsilon}} > -\lambda_\gamma (\dfrac{\partial \mathcal{L}_c}{\partial \hat{h}_c})^{-1}, & \gamma_c > 0 \\ \dfrac{(h_{z/x,c} - \mu_{z/x,c})}{\sqrt{\sigma_{z/x,c}^2 + \epsilon}} < \lambda_\gamma (\dfrac{\partial \mathcal{L}_c}{\partial \hat{h}_c})^{-1}, & \gamma_c < 0 \end{cases} \tag{7.13}$$

With the central limit theorem, we can convert the above inequality into the probability formulation:

$$\mathbb{P}\Big(-\lambda_\gamma(\frac{\partial \mathcal{L}_c}{\partial \hat{h}_c})^{-1} < \frac{(h_{z/x,c} - \mu_{z/x,c})}{\sqrt{\sigma^2_{z/x,c} + \epsilon}} < \lambda_\gamma(\frac{\partial \mathcal{L}_c}{\partial \hat{h}_c})^{-1}\Big) = 2\Phi\big(\lambda_\gamma(\frac{\partial \mathcal{L}_c}{\partial h_c})^{-1}\big) - 1. \tag{7.14}$$

The model convergence means $\frac{\partial \mathcal{L}_c}{\partial \hat{h}_c} \to 0$ so that the above probability approximates one. It suggests the scaling factors to these channels will become zero with high-probability. Multi-view images are likely to activate different channels in this part for the classification task. Thus, we consider these channel information as view-specific content. Inversely, benefit from the ℓ_2-norm analysis [181], the $\gamma_c, c \in [1, C/2)$ will be dense non-zero values with the constraint $\sum_{c=1}^{\lfloor C/2 \rfloor} \gamma^2_{l,c}$. These channels across various views are both activated to learn semantic from the identical location of images or feature maps and tend to include the similar even consistent patterns, which are defined as view-common channels.

7.3 Experiments

7.3.1 Experimental Details

Datasets: i). **RGB-D** dataset [182] is a large-scale household objects dataset including 51 categories and each specific object is captured by Kinect style 3D camera (30Hz) generating RGB and depth images at the same time. ii). **B3DO** [183] is a popular 3D benchmark database with RGB and depth image pairs from 83 object categories. And these images are collected from real domestic and office-environments by Microsoft Kinect sensor. iii). **Office-31** [85] is a standard multi-domain RGB image benchmark including Amazon (**A**), Webcam (**W**) and DSLR (**D**), which are gathered with different cameras. And all domains share the identical label space with 31 categories. iv). **Office-Home** [86] as a large-scale cross-domain dataset involves four domains as Art Painting (**Ar**), Clipart (**Cl**), Product (**Pr**) and Real World (**Rw**) with significant image style difference. And each domain includes the same 65 object classes. v). **Caltech-256** (**C**) [184] is a classical natural image database with 30,607 images from 257 objects.

In IMVDA experiments, we consider RGB-D and B3DO as two multi-view (RGB and Depth) well-annotated source domains, while the Caltech-256 or each domain of Office-

Table 7.1: Object Classification Accuracy (%) of target domain with **RGB-D** datasets as multi-view source domain. We adopt **bold** to highlight the best result and show the second best one with underline.

Method	A	D	W	Ar	Cl	Pr	Rw	C	Avg
ResNet [76]	61.75	79.37	81.73	35.90	28.86	48.01	52.68	74.82	57.89
DANN [36]	67.98	81.51	82.35	46.42	35.50	48.99	63.15	75.42	62.67
CDAN+E [19]	66.15	84.37	85.06	46.95	34.42	51.04	63.30	78.32	63.70
SRDC [20]	68.28	87.70	87.77	51.57	35.96	58.00	66.44	81.45	67.14
CGDM [185]	65.48	84.57	84.59	43.26	36.80	53.54	63.20	77.49	63.62
FixBi [186]	69.07	85.04	86.59	50.29	**38.33**	61.53	65.58	81.14	67.19
M3SDA [68]	66.11	85.70	85.86	45.10	37.00	56.53	64.96	80.71	65.25
DRT [187]	67.86	86.79	86.57	46.00	35.55	57.28	64.97	80.62	65.71
Ours	**70.79**	**89.68**	**90.87**	**56.17**	35.46	**66.86**	**70.33**	**84.21**	**70.55**

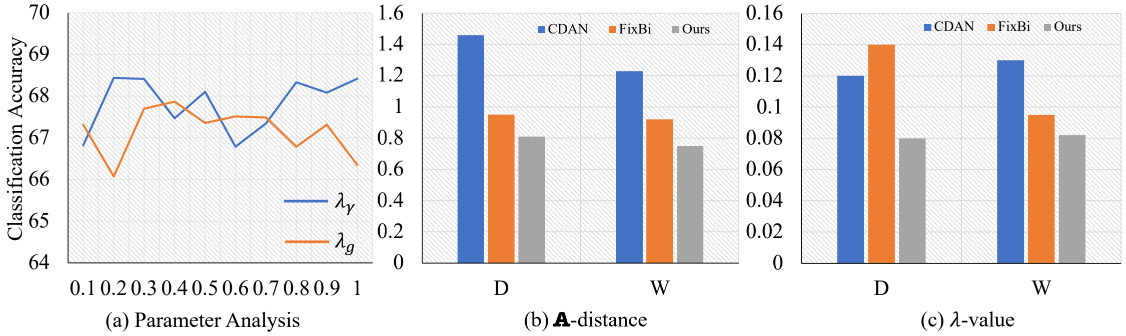
31 and Office-Home as the unlabeled target domain to mimic the incomplete multi-view scenario. For each specific adaptation task, we select the shared categories across source and target domains. Concretely, the number of categories for tasks **RGB-D→Office31**, **RGB-D→Office-Home**, **RGB-D→Caltech-256** are 8, 13 and 10, respectively, while that for **B3DO→Office31**, **B3DO→Office-Home**, **B3DO→Caltech-256** are 27, 14 and 8, respectively.

Implementation Details: The implementation of our model is based on pytorch platform. And we adopt the pre-trained ResNet-50 [76] without the last FC layer as the feature extractor for source and target models, and $W_{s/t} \in \mathbb{R}^{64\times 256}$, $\{\mathbf{F}_i^{x/z}, \mathbf{F}_i^{xz}, \mathbf{G}_i^{s/t}, \bar{\mathbf{G}}_i^{s/t}\} \in \mathbb{R}^{256}$. Moreover, the stochastic gradient descent (SGD) optimizer with momentum 0.9 is used to optimize all parameters. The learning rate and batch size are 1e-3 and 96. The ϵ and δ are set as 1e-6 and 0.02 for all experiments.

Baselines: In term of IMVDA, since source and target domains both involve one identical view data, the conventional unsupervised domain adaptation methods can exploit these samples to achieve alignment and identify target samples. Thus, we evaluate the DANN [36], CDAN+E [19], SRDC [20], CGDM [185], FixBi [186] under IMVDA scenario. Moreover, each view data of source domain can be considered as one independent domain. The multi-source domain adaptation methods M3SDA [68] and DRT [187] are used to solve IMVDA challenges. And we adopt their published source code and empirically search optimal parameters to conduct experiments.

Table 7.2: Object Classification Accuracy (%) of target domain with **B3DO** datasets as multi-view source domain. We adopt **bold** to highlight the best result and show the second best one with underline.

Method	A	D	W	Ar	Cl	Pr	Rw	C	Avg
ResNet [76]	31.98	49.54	44.35	48.54	35.53	50.56	57.70	48.56	45.85
DANN [36]	44.05	63.53	62.35	59.61	40.05	67.09	74.98	68.18	59.98
CDAN+E [19]	47.70	66.75	64.69	62.00	43.93	70.29	77.93	71.35	63.08
SRDC [20]	49.47	68.67	66.74	64.44	45.85	72.77	79.73	73.55	65.15
CGDM [185]	47.19	66.07	64.09	61.23	43.15	69.70	76.97	70.42	62.35
FixBi [186]	49.67	68.59	66.69	63.97	45.41	71.72	80.23	72.82	64.89
M3SDA [68]	47.76	66.55	64.92	62.01	44.86	71.17	77.51	71.96	63.34
DRT [187]	47.75	67.59	66.01	63.00	44.22	70.84	78.62	72.82	63.86
Ours	**50.02**	**71.87**	**70.23**	**68.00**	**47.40**	**76.61**	**82.81**	**77.21**	**68.02**



(a) Parameter Analysis

(b) $\mathcal{A}$-distance

(c) λ-value

Figure 7.2: Parameter analysis & Transfer ability. (a) Target classification accuracy with the varying parameters λ_γ and λ_g from 0.1 to 1.0 with B3DO as source domain. (b) $\mathcal{A}$-distance of source and target features from the same view data with RGB-D as source domain. (c) λ-value of three methods with tasks from RGB-D to D and W.

7.3.2 Comparison of Results

The main experimental results in terms of target recognition accuracy are summarized in Table 7.1 and Table 7.2. According to the evaluation performance, we can easily achieve several significant conclusions. **First**, our method outperforms other baselines by a large margin on the average classification accuracy. Specifically, with RGB-D dataset as source domain, our CEKT surpasses the second best comparison (i.e., FixBi) by 3.36%. It illustrates the deployment of multi-view information effectively boosts the model performance on target domain even with considerable distribution shift. **Second**, we notice that our CEKT obtains much higher classification accuracy than others on the task RGB-D→Ar. As we all know, the images of Art Painting domain in Office-Home include lots of texture information to describe each object. On the other hand, depth sensor integrates more spatial information into depth images to clearly show the contour of object, which provides more discriminative semantic to the classification task. However, M3SDA and DRT, taking advantage of depth images to train the model, still fail to effectively assist the recognition of unlabeled target samples. These observations demonstrate our proposed solution not only emphasizes the specific semantic of depth images via source cross-view channel enhancement but also transfers such knowledge from source domain to target domain by reducing the negative influence of missing view with adaptive knowledge transfer network. **Third**, comparison of Table 7.1 and Table 7.2 shows that B3DO has larger distribution difference than RGB-D to the other target domain in Office-31, Office-Home and Caltech-256, as we achieve worse results by directly recognizing the target based on ResNet features. However, our proposed CEKT model can still achieve very close results no matter which source is used. In details, we improve the average accuracy from 57.89% to 70.55% by using RGB-D as source, while promote the average accuracy from 45.85% to 68.02% by using B3DO as source.

7.3.3 Empirical Analysis

Parameter Sensitivity. During training model, there are two parameters (λ_γ, λ_g) in our designed CEKT framework which are manually adjusted. These two parameters are changed from 0.1 to 1.0 with step size 0.1. To analyse the model sensitivity to them, we record the classification accuracy of target domain with various parameter selection on task from B3DO to **Ar**, which is shown in Figure 7.2 (a). On the whole, the model is not sensitive to the change of parameters. However, larger λ_γ can easily bring more benefits to the model, while the smaller λ_g results in better performance, which further illustrates the proposed channel enhancement module effectively assists model to learning discriminative features. Note that for the selection of parameters, we randomly select 10% source samples as validation set for each tentative and use it to evaluate the model performance.

Transfer Ability. In addition, Ben-David theoretically points out the learning bound of domain adaptation [188] is determined by three parts: 1) the expected error $\varepsilon_s(h)$ of hypothesis h on source domain; 2) the $\mathcal{A}$-distance defined as $d_{\mathcal{H}\triangle\mathcal{H}} = 2(1-2\xi)$ measuring the domain mismatch, where ξ is the error from a trained domain classifier distinguishing source from target ones; 3) the error λ produced by the ideal hypothesis on both two

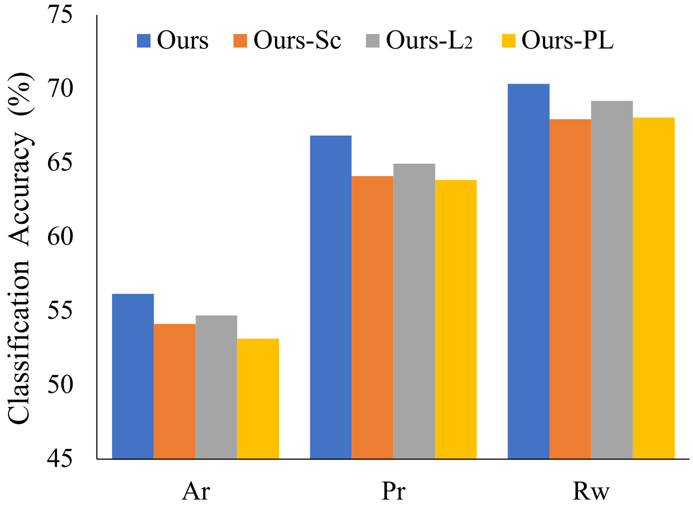
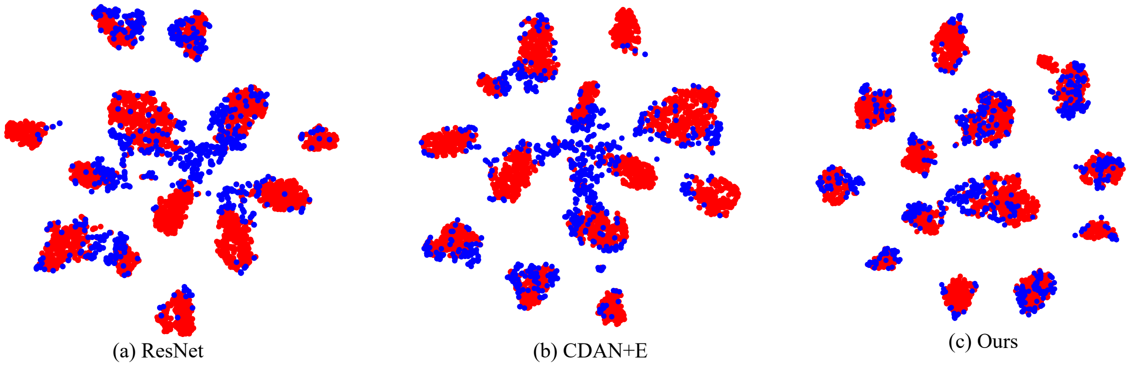


Figure 7.3: Ablation study of model variants on three tasks with RGB-D as source domain.

domains. Inspired by this theoretical analysis, we report the $\mathcal{A}$-distance and λ-value over the shared-view data across source and target domains and show the results in Figure 7.2 (b)-(c). Compared with CDAN and FixBi, our proposed method obtains relative smaller $\mathcal{A}$-distance and λ-value on two tasks from RGB-D to **D** and **W**, which suggests that CEKT learns a model with a higher generalization ability.

Ablation Study. To clearly reflect the contribution of each component to the model performance, we carry out experiments on three knowledge transfer tasks with RGB-D as source domain by removing the corresponding operations. As previous mentioned, the source channel enhanced network actively discovers the view-common and view-specific parts via Eq. (7.3) and encourages the representation of important channels with Eq. (7.9). Thus, we replace Eq. (7.3) with $\sum_{c=1}^{C} |\gamma_{x/z,c}|$ (Ours-L$_2$) and attempt to remove Eq. (7.9) as Ours-Sc to study their effect. In addition, the model training adopts pseudo labels to facilitate feature with more discriminative power, and we further add a variant without the pseudo labeling as Ours-PL. Figure 7.3 reports the corresponding results with various methods on three tasks. According to it, we discover the enhancement with channel similarity and pseudo labels both produce significant and positive influence on improving model performance on target domain. Moreover, the sparse constraint for parameters $\gamma_{x/z,c}$ as [177] also results in the performance degradation, which further verifies the necessity of the preservation for the view-common channel split in multi-view data analysis.



(a) ResNet

(b) CDAN+E

(c) Ours

Figure 7.4: Feature Visualization with t-SNE in 2D plane. The source and target features are represented by red and blue, respectively. And the experiment aims to transfer knowledge from RGB-D to **Ar** in Office-Home.

Feature Visualization. To further understand the situation of distribution alignment, we follow [42] to visualize source and target features from the same view in 2D-plane, shown in Figure 7.4. Concretely, we access to the high-level features $\bar{\mathbf{G}}_i^{s/t}$ from the well-trained target model and adopt t-SNE technique to draw them in the canvas. Moreover, the experiment is carried out on adaptation task from RGB-D to **Pr** and ResNet as well as CDAN+E are considered as the competitors. According to the visualization results, it is easy to observe that there exist more overlaps between source and target features, compared with other baselines, which shows our method successfully mitigates the domain shift and better align them. Moreover, we notice that the classification boundary is more explicit than that in ResNet and CDAN+E. It suggests CEKT effectively learns the discriminative features for classification task.

7.4 Conclusion

Unsupervised domain adaptation (UDA) aims to learn the domain-invariant knowledge across well-supervised source and unlabeled target samples to enhance the model generalization ability. However, UDA assumes the instances per domain are captured by single sensor, which difficultly matches the practical scenario with multi-view data. This paper considered a practical and challenging problem named incomplete multi-view domain adaptation (IMVDA) which access to multi-view source data and single-view target samples. To overcome the challenge, we proposed a novel learning framework channel enhancement and knowledge transfer (CEKT). Concretely, CEKT first explored channel attributions to conduct semantic fusion and enhance the representation of view-common channels to learn more discriminative features. Moreover, adaptive knowledge transfer module not only brought multi-view knowledge to single-view feature learning but also achieved simple yet effective alignment across source and target domains. Considerable experimental results and analysis fully demonstrated our CEKT effectively broke the bottleneck of IMVDA by improving the performance.

Chapter 8

Conclusion & Future Works

Transfer learning is the lubricant that drives widespread adoption of deep neural network in industrial application by mitigating distribution shift. One representative problem setting of transfer learning is unsupervised domain adaptation (UDA) leveraging well-annotated source domain and unlabeled target samples to conduct knowledge transfer and adaptation. But real-world application scenarios difficultly guarantee that the model training can access all domain data due to privacy concern and inconvenient data collection. In this thesis, we mainly explore domain adaptation with multiple data access privileges, e.g., source-free domain adaptation, target-data absent domain adaptation and incomplete multi-view domain adaptation.

In Chapters 2 & 3, we analyze the limitation of existing UDA methods on achieving distribution alignment and explore intrinsic cross-domain structural information to instruct the domain-invariant feature learning. Concretely, we utilize the structural knowledge to build an intermediate domain and regard it as the bridge to connect source and target domains via metric learning and adversarial mechanism.

In Chapter 4, we mainly consider source-free domain adaptation problem, where only the well-trained source model and unlabeled target samples are allowed to participant in the process of knowledge adaption. To reduce the negative effect of source-data absence, our proposed method adopts dual-classifier to distinguish source-similar samples from source-dissimilar ones and utilizes adversarial and contrastive constraints to gradually align them.

In Chapters 5 & 6, we take the target-data absent scenario into account, where the entire model training process fails to access samples collected from the unseen target distribution. For this problem, we explore two different learning strategies. One is generating domain-invariant representation over multiple source domains to promote the generalization of the model. The other one instructs the model to decompose features into task-relevant semantics and task-irrelevant ones by training it on an additional multi-modality dataset. Under this condition, the model can further extract and generalize intrinsic source knowledge.

In Chapter 7, we explore incomplete multi-view domain adaption where source samples are captured by multiple views (sensors) while target instances are collected by one single view. The main challenge is how to effectively transfer sufficient multi-view semantic information to benefit the task on the target domain. For this, we adopt a channel-wise exchange mechanism and optimal transport to fulfill representation fusion and knowledge transfer.

In future works, we will follow the current research direction and continuously lift the limitation of applying transfer learning techniques by considering more practical industrial scenarios with constraints on data accessibility. The potential research topics are federated domain adaptation and personalized federated learning. In the former one, source and target data are stored in different clients and only used to train their local models without cross-domain data exchange. But they can communicate with the third-party server to achieve knowledge transfer and benefit target tasks by uploading and downloading network parameters. Differently, the latter expects to utilize knowledge sharing among various clients to further improve the performance of all local models. To increase the compatibility of knowledge, eliminating domain-specific semantics and learning domain-invariant features are the core technical demand for overcoming this problem.

Bibliography

[1] A. Floris and L. Atzori, “Quality of experience in the multimedia internet of things: Definition and practical use-cases,” in *2015 IEEE International Conference on Communication Workshop (ICCW)*, pp. 1747–1752, IEEE, 2015.

[2] A. Nauman, Y. A. Qadri, M. Amjad, Y. B. Zikria, M. K. Afzal, and S. W. Kim, “Multimedia internet of things: A comprehensive survey,” *IEEE Access*, vol. 8, pp. 8202–8250, 2020.

[3] B. S. Alhayani *et al.*, “Retracted article: Visual sensor intelligent module based image transmission in industrial manufacturing for monitoring and manipulation problems,” *Journal of Intelligent Manufacturing*, vol. 32, no. 2, pp. 597–610, 2021.

[4] G. Chen, H. Cao, J. Conradt, H. Tang, F. Rohrbein, and A. Knoll, “Event-based neuromorphic vision for autonomous driving: A paradigm shift for bio-inspired visual sensing and perception,” *IEEE Signal Processing Magazine*, vol. 37, no. 4, pp. 34–49, 2020.

[5] S.-h. Zhong, Y. Liu, and Y. Liu, “Bilinear deep learning for image classification,” in *Proceedings of the 19th ACM international conference on Multimedia*, pp. 343–352, 2011.

[6] H. Cevikalp, B. Benligiray, and O. N. Gerek, “Semi-supervised robust deep neural networks for multi-label image classification,” *Pattern Recognition*, vol. 100, p. 107164, 2020.

[7] D. Erhan, C. Szegedy, A. Toshev, and D. Anguelov, “Scalable object detection using deep neural networks,” in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 2147–2154, 2014.

[8] Z.-Q. Zhao, P. Zheng, S.-t. Xu, and X. Wu, “Object detection with deep learning: A review,” *IEEE transactions on neural networks and learning systems*, vol. 30, no. 11, pp. 3212–3232, 2019.

[9] J. Lee, E. Kim, S. Lee, J. Lee, and S. Yoon, “Ficklenet: Weakly and semi-supervised semantic image segmentation using stochastic inference,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 5267–5276, 2019.

[10] X. Luo, J. Chen, T. Song, and G. Wang, “Semi-supervised medical image segmentation through dual-task consistency,” in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 35, pp. 8801–8809, 2021.

[11] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "Imagenet: A large-scale hierarchical image database," in *2009 IEEE conference on computer vision and pattern recognition*, pp. 248–255, Ieee, 2009.

[12] D. Hoiem, S. K. Divvala, and J. H. Hays, "Pascal voc 2008 challenge," *World Literature Today*, vol. 24, 2009.

[13] T. Jing, H. Xia, R. Tian, H. Ding, X. Luo, J. Domeyer, R. Sherony, and Z. Ding, "Inaction: Interpretable action decision making for autonomous driving," in *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XXXVIII*, pp. 370–387, Springer, 2022.

[14] J. Huang, D. Guan, A. Xiao, S. Lu, and L. Shao, "Category contrast for unsupervised domain adaptation in visual tasks," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 1203–1214, 2022.

[15] H. Xia and Z. Ding, "Cross-domain collaborative normalization via structural knowledge," in *AAAI 2022*, 2022.

[16] J. Lee and G. Lee, "Feature alignment by uncertainty and self-training for source-free unsupervised domain adaptation," *Neural Networks*, vol. 161, pp. 682–692, 2023.

[17] H. Xia, T. Jing, and Z. Ding, "Generative inference network for imbalanced domain generalization," *IEEE Transactions on Image Processing*, 2023.

[18] S. Ben-David, J. Blitzer, K. Crammer, A. Kulesza, F. Pereira, and J. W. Vaughan, "A theory of learning from different domains," *Machine learning*, vol. 79, no. 1-2, pp. 151–175, 2010.

[19] M. Long, Z. Cao, J. Wang, and M. I. Jordan, "Conditional adversarial domain adaptation," *arXiv preprint arXiv:1705.10667*, 2017.

[20] H. Tang, K. Chen, and K. Jia, "Unsupervised domain adaptation via structurally regularized deep clustering," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 8725–8735, 2020.

[21] M. Long, Y. Cao, J. Wang, and M. I. Jordan, "Learning transferable features with deep adaptation networks," *arXiv preprint arXiv:1502.02791*, 2015.

[22] Y. Zhang, H. Tang, K. Jia, and M. Tan, "Domain-symmetric networks for adversarial domain adaptation," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 5031–5040, 2019.

[23] F. Lv, J. Liang, S. Li, B. Zang, C. H. Liu, Z. Wang, and D. Liu, "Causality inspired representation learning for domain generalization," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 8046–8056, 2022.

[24] J. Wang, C. Lan, C. Liu, Y. Ouyang, T. Qin, W. Lu, Y. Chen, W. Zeng, and P. Yu, "Generalizing to unseen domains: A survey on domain generalization," *IEEE Transactions on Knowledge and Data Engineering*, 2022.

[25] Z. Ding and Y. Fu, "Low-rank common subspace for multi-view learning," in *IEEE International Conference on Data Mining*, 2014.

[26] Y. He, Y. Tian, and D. Liu, "Multi-view transfer learning with privileged learning framework," *Neurocomputing*, vol. 335, pp. 131–142, 2019.

[27] M. Long, H. Zhu, J. Wang, and M. I. Jordan, "Deep transfer learning with joint adaptation networks," in *Proceedings of the 34th International Conference on Machine Learning-Volume 70*, pp. 2208–2217, JMLR. org, 2017.

[28] B. Sun and K. Saenko, "Deep coral: Correlation alignment for deep domain adaptation," in *European Conference on Computer Vision Workshops*, pp. 443–450, Springer, 2016.

[29] W. Zellinger, T. Grubinger, E. Lughofer, T. Natschläger, and S. Saminger-Platz, "Central moment discrepancy (cmd) for domain-invariant representation learning," *arXiv preprint arXiv:1702.08811*, 2017.

[30] G. Kang, L. Jiang, Y. Yang, and A. G. Hauptmann, "Contrastive adaptation network for unsupervised domain adaptation," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 4893–4902, 2019.

[31] Y. Zhang, B. Deng, K. Jia, and L. Zhang, "Label propagation with augmented anchors: A simple semi-supervised learning baseline for unsupervised domain adaptation," in *European Conference on Computer Vision*, pp. 781–797, Springer, 2020.

[32] H. Yan, Y. Ding, P. Li, Q. Wang, Y. Xu, and W. 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*, pp. 2272–2281, 2017.

[33] Y. Pan, T. Yao, Y. Li, C.-W. Ngo, and T. Mei, "Exploring category-agnostic clusters for open-set domain adaptation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 13867–13875, 2020.

[34] M. Li, Y.-M. Zhai, Y.-W. Luo, P.-F. Ge, and C.-X. Ren, "Enhanced transport distance for unsupervised domain adaptation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 13936–13944, 2020.

[35] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial nets," in *Proceedings of Advances in Neural Information Processing Systems*, pp. 2672–2680, 2014.

[36] Y. Ganin, E. Ustinova, H. Ajakan, P. Germain, H. Larochelle, F. Laviolette, M. Marchand, and V. Lempitsky, "Domain-adversarial training of neural networks," *The Journal of Machine Learning Research*, vol. 17, no. 1, pp. 2096–2030, 2016.

[37] M. Mirza and S. Osindero, "Conditional generative adversarial nets," *arXiv preprint arXiv:1411.1784*, 2014.

[38] M. Long, Z. Cao, J. Wang, and M. I. Jordan, "Conditional adversarial domain adaptation," in *Advances in Neural Information Processing Systems*, pp. 1640–1650, 2018.

[39] H. Tang and K. Jia, "Discriminative adversarial domain adaptation.," in *Proceeding of the AAAI Conference on Artificial Intelligence*, pp. 5940–5947, 2020.

[40] J. Hoffman, E. Tzeng, T. Park, J.-Y. Zhu, P. Isola, K. Saenko, A. Efros, and T. Darrell, "Cycada: Cycle-consistent adversarial domain adaptation," in *International conference on machine learning*, pp. 1989–1998, PMLR, 2018.

[41] M. Xu, J. Zhang, B. Ni, T. Li, C. Wang, Q. Tian, and W. Zhang, "Adversarial domain adaptation with domain mixup," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 34, pp. 6502–6509, 2020.

[42] R. Li, Q. Jiao, W. Cao, H.-S. Wong, and S. Wu, "Model adaptation: Unsupervised domain adaptation without source data," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 9641–9650, 2020.

[43] V. K. Kurmi, V. K. Subramanian, and V. P. Namboodiri, "Domain impression: A source data free domain adaptation method," in *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pp. 615–625, 2021.

[44] J. Liang, D. Hu, and J. Feng, "Do we really need to access the source data? source hypothesis transfer for unsupervised domain adaptation," in *International Conference on Machine Learning*, pp. 6028–6039, PMLR, 2020.

[45] S. Lin, C.-T. Li, and A. C. Kot, "Multi-domain adversarial feature generalization for person re-identification," *IEEE Transactions on Image Processing*, vol. 30, pp. 1596–1607, 2020.

[46] Z. Ding and Y. Fu, "Deep domain generalization with structured low-rank constraint," *IEEE Transactions on Image Processing*, vol. 27, no. 1, pp. 304–313, 2017.

[47] X. Yue, Y. Zhang, S. Zhao, A. Sangiovanni-Vincentelli, K. Keutzer, and B. Gong, "Domain randomization and pyramid consistency: Simulation-to-real generalization without accessing target domain data," in *Proceedings of the IEEE International Conference on Computer Vision*, pp. 2100–2110, 2019.

[48] S. Zakharov, W. Kehl, and S. Ilic, "Deceptionnet: Network-driven domain randomization," in *Proceedings of the IEEE International Conference on Computer Vision*, pp. 532–541, 2019.

[49] Q. Xu, R. Zhang, Y. Zhang, Y. Wang, and Q. Tian, "A fourier-based framework for domain generalization," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 14383–14392, 2021.

[50] Y. Li, X. Tian, M. Gong, Y. Liu, T. Liu, K. Zhang, and D. Tao, "Deep domain generalization via conditional invariant adversarial networks," in *Proceedings of the European Conference on Computer Vision*, pp. 624–639, 2018.

[51] D. Li, Y. Yang, Y.-Z. Song, and T. M. Hospedales, "Deeper, broader and artier domain generalization," in *Proceedings of the IEEE International Conference on Computer Vision*, pp. 5542–5550, 2017.

[52] D. Li, J. Zhang, Y. Yang, C. Liu, Y.-Z. Song, and T. M. Hospedales, "Episodic training for domain generalization," in *Proceedings of the IEEE International Conference on Computer Vision*, pp. 1446–1455, 2019.

[53] S. Wang, L. Yu, C. Li, C.-W. Fu, and P.-A. Heng, "Learning from extrinsic and intrinsic supervisions for domain generalization," in *Proceedings of the European Conference on Computer Vision*, pp. 159–176, Springer, 2020.

[54] F. M. Carlucci, A. D'Innocente, S. Bucci, B. Caputo, and T. Tommasi, "Domain generalization by solving jigsaw puzzles," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 2229–2238, 2019.

[55] A. Gretton, A. Smola, J. Huang, M. Schmittfull, K. Borgwardt, and B. Schölkopf, "Covariate shift by kernel mean matching," *Dataset shift in machine learning*, vol. 3, no. 4, p. 5, 2009.

[56] Z. Chen, J. Zhuang, X. Liang, and L. Lin, "Blending-target domain adaptation by adversarial meta-adaptation networks," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 2248–2257, 2019.

[57] S. Jiang, Z. Ding, and Y. Fu, "Heterogeneous recommendation via deep low-rank sparse collective factorization," *IEEE transactions on pattern analysis and machine intelligence*, 2019.

[58] Z. Ding, M. Shao, and Y. Fu, "Robust multi-view representation: a unified perspective from multi-view learning to domain adaption," in *Proceedings of the 27th International Joint Conference on Artificial Intelligence*, pp. 5434–5440, 2018.

[59] J. Dong, Y. Cong, G. Sun, B. Zhong, and X. Xu, "What can be transferred: Unsupervised domain adaptation for endoscopic lesions segmentation," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2020.

[60] G. Yang, H. Xia, M. Ding, and Z. Ding, "Bi-directional generation for unsupervised domain adaptation," in *Thirty-Fourth AAAI Conference on Artificial Intelligence*, 2020.

[61] S. Li, C. H. Liu, Q. Lin, Q. Wen, L. Su, G. Huang, and Z. Ding, "Deep residual correction network for partial domain adaptation," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2020.

[62] S. Roy, A. Siarohin, E. Sangineto, S. R. Bulo, N. Sebe, and E. Ricci, "Unsupervised domain adaptation using feature-whitening and consensus loss," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 9471–9480, 2019.

[63] Z. Ding, S. Li, M. Shao, and Y. Fu, "Graph adaptive knowledge transfer for unsupervised domain adaptation," in *Proceedings of the European Conference on Computer Vision*, pp. 37–52, 2018.

[64] H. Liu, M. Long, J. Wang, and M. Jordan, "Transferable adversarial training: A general approach to adapting deep classifiers," in *International Conference on Machine Learning*, pp. 4013–4022, 2019.

[65] Y. Luo, L. Zheng, T. Guan, J. Yu, and Y. Yang, "Taking a closer look at domain shift: Category-level adversaries for semantics consistent domain adaptation," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 2507–2516, 2019.

[66] W. Zhang, W. Ouyang, W. Li, and D. Xu, "Collaborative and adversarial network for unsupervised domain adaptation," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 3801–3809, 2018.

[67] S. Xie, Z. Zheng, L. Chen, and C. Chen, "Learning semantic representations for unsupervised domain adaptation," in *International Conference on Machine Learning*, pp. 5419–5428, 2018.

[68] X. Peng, Q. Bai, X. Xia, Z. Huang, K. Saenko, and B. Wang, "Moment matching for multi-source domain adaptation," in *Proceedings of the IEEE International Conference on Computer Vision*, pp. 1406–1415, 2019.

[69] A. Kumar, P. Sattigeri, K. Wadhawan, L. Karlinsky, R. Feris, B. Freeman, and G. Wornell, "Co-regularized alignment for unsupervised domain adaptation," in *Advances in Neural Information Processing Systems*, pp. 9345–9356, 2018.

[70] K. Saito, K. Watanabe, Y. Ushiku, and T. Harada, "Maximum classifier discrepancy for unsupervised domain adaptation," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 3723–3732, 2018.

[71] C.-Y. Lee, T. Batra, M. H. Baig, and D. Ulbricht, "Sliced wasserstein discrepancy for unsupervised domain adaptation," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 10285–10295, 2019.

[72] H. Xu, D. Luo, H. Zha, and L. Carin, "Gromov-wasserstein learning for graph matching and node embedding," *arXiv preprint arXiv:1901.06003*, 2019.

[73] S. Lee, D. Kim, N. Kim, and S.-G. Jeong, "Drop to adapt: Learning discriminative features for unsupervised domain adaptation," in *Proceedings of the IEEE International Conference on Computer Vision*, pp. 91–100, 2019.

[74] S. Cicek and S. Soatto, "Unsupervised domain adaptation via regularized conditional alignment," *arXiv preprint arXiv:1905.10885*, 2019.

[75] S. Song, S. P. Lichtenberg, and J. Xiao, "Sun rgb-d: A rgb-d scene understanding benchmark suite," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 567–576, 2015.

[76] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 770–778, 2016.

[77] T. Vayer, L. Chapel, R. Flamary, R. Tavenard, and N. Courty, "Fused gromov-wasserstein distance for structured objects: theoretical foundations and mathematical properties," *arXiv preprint arXiv:1811.02834*, 2018.

[78] G. Peyré, M. Cuturi, and J. Solomon, "Gromov-wasserstein averaging of kernel and distance matrices," in *International Conference on Machine Learning*, pp. 2664–2672, 2016.

[79] Y. Zhang, P. David, and B. Gong, "Curriculum domain adaptation for semantic segmentation of urban scenes," in *Proceedings of the IEEE International Conference on Computer Vision*, pp. 2020–2030, 2017.

[80] T.-H. Vu, H. Jain, M. Bucher, M. Cord, and P. Pérez, "Advent: Adversarial entropy minimization for domain adaptation in semantic segmentation," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 2517–2526, 2019.

[81] P. O. Pinheiro, "Unsupervised domain adaptation with similarity learning," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 8004–8013, 2018.

[82] X. Wang, L. Li, W. Ye, M. Long, and J. Wang, "Transferable attention for domain adaptation," in *AAAI Conference on Artificial Intelligence (AAAI)*, 2019.

[83] R. Xu, G. Li, J. Yang, and L. Lin, "Larger norm more transferable: An adaptive feature norm approach for unsupervised domain adaptation," in *Proceedings of the IEEE International Conference on Computer Vision*, pp. 1426–1435, 2019.

[84] R. Cai, Z. Li, P. Wei, J. Qiao, K. Zhang, and Z. Hao, "Learning disentangled semantic representation for domain adaptation," in *IJCAI: proceedings of the conference*, vol. 2019, p. 2060, NIH Public Access, 2019.

[85] K. Saenko, B. Kulis, M. Fritz, and T. Darrell, "Adapting visual category models to new domains," in *European conference on computer vision*, pp. 213–226, Springer, 2010.

[86] H. Venkateswara, J. Eusebio, S. Chakraborty, and S. Panchanathan, "Deep hashing network for unsupervised domain adaptation," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 5018–5027, 2017.

[87] Y. Pei, Y. Huang, Q. Zou, X. Zhang, and S. Wang, "Effects of image degradation and degradation removal to cnn-based image classification," *IEEE transactions on pattern analysis and machine intelligence*, 2019.

[88] H. Wu, Y. Chen, N. Wang, and Z. Zhang, "Sequence level semantics aggregation for video object detection," in *Proceedings of the IEEE International Conference on Computer Vision*, pp. 9217–9225, 2019.

[89] C. Liu, L.-C. Chen, F. Schroff, H. Adam, W. Hua, A. L. Yuille, and L. Fei-Fei, "Auto-deeplab: Hierarchical neural architecture search for semantic image segmentation," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 82–92, 2019.

[90] F. Wang, M. Jiang, C. Qian, S. Yang, C. Li, H. Zhang, X. Wang, and X. Tang, "Residual attention network for image classification," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 3156–3164, 2017.

[91] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *arXiv preprint arXiv:1409.1556*, 2014.

[92] H. Zhao, S. Zhang, G. Wu, J. M. Moura, J. P. Costeira, and G. J. Gordon, "Adversarial multiple source domain adaptation," in *Advances in neural information processing systems*, pp. 8559–8570, 2018.

[93] Y. Li, N. Wang, J. Shi, X. Hou, and J. Liu, "Adaptive batch normalization for practical domain adaptation," *Pattern Recognition*, vol. 80, pp. 109–117, 2018.

[94] Y. Pan, T. Yao, Y. Li, Y. Wang, C.-W. Ngo, and T. Mei, "Transferrable prototypical networks for unsupervised domain adaptation," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 2239–2247, 2019.

[95] J. Liang, R. He, Z. Sun, and T. Tan, "Exploring uncertainty in pseudo-label guided unsupervised domain adaptation," *Pattern Recognition*, vol. 96, p. 106996, 2019.

[96] H. Xia and Z. Ding, "Structure preserving generative cross-domain learning," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 4364–4373, 2020.

[97] A. Iyer, S. Nath, and S. Sarawagi, "Maximum mean discrepancy for class ratio estimation: Convergence bounds and kernel selection," in *International Conference on Machine Learning*, pp. 530–538, 2014.

[98] P. Rodríguez, I. Laradji, A. Drouin, and A. Lacoste, "Embedding propagation: Smoother manifold for few-shot classification," in *European Conference on Computer Vision*, pp. 121–138, Springer, 2020.

[99] D. Zhou, O. Bousquet, T. Lal, J. Weston, and B. Schölkopf, "Learning with local and global consistency," *Advances in neural information processing systems*, vol. 16, 2003.

[100] M. Chen, S. Zhao, H. Liu, and D. Cai, "Adversarial-learned loss for domain adaptation.," in *Proceeding of the AAAI Conference on Artificial Intelligence*, pp. 3521–3528, 2020.

[101] Y.-W. Luo, C.-X. Ren, D. Dao-Qing, and H. Yan, "Unsupervised domain adaptation via discriminative manifold propagation," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2020.

[102] S. Cui, S. Wang, J. Zhuo, C. Su, Q. Huang, and Q. Tian, "Gradually vanishing bridge for adversarial domain adaptation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 12455–12464, 2020.

[103] Y. Wu, D. Inkpen, and A. El-Roby, "Dual mixup regularized learning for adversarial domain adaptation," in *European Conference on Computer Vision*, pp. 540–555, Springer, 2020.

[104] X. Peng, B. Usman, N. Kaushik, J. Hoffman, D. Wang, and K. Saenko, "Visda: The visual domain adaptation challenge," *arXiv preprint arXiv:1710.06924*, 2017.

[105] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick, "Microsoft coco: Common objects in context," in *European conference on computer vision*, pp. 740–755, Springer, 2014.

[106] S. Li, B. Xie, Q. Lin, C. H. Liu, G. Huang, and G. Wang, "Generalized domain conditioned adaptation network," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2021.

[107] E. Tzeng, J. Hoffman, K. Saenko, and T. Darrell, "Adversarial discriminative domain adaptation," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 7167–7176, 2017.

[108] S. Li, C. Liu, Q. Lin, B. Xie, Z. Ding, G. Huang, and J. Tang, "Domain conditioned adaptation network," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 34, pp. 11386–11393, 2020.

[109] G. French, M. Mackiewicz, and M. Fisher, "Self-ensembling for visual domain adaptation," *arXiv preprint arXiv:1706.05208*, 2017.

[110] J. Liang, N. Homayounfar, W.-C. Ma, Y. Xiong, R. Hu, and R. Urtasun, "Polytransform: Deep polygon transformer for instance segmentation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 9131–9140, 2020.

[111] H. Zhao, Z. Ding, and Y. Fu, "Multi-view clustering via deep matrix factorization," in *Thirty-First AAAI Conference on Artificial Intelligence*, 2017.

[112] Z. Liu, G. Gao, L. Sun, and L. Fang, "Ipg-net: Image pyramid guidance network for small object detection," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pp. 1026–1027, 2020.

[113] H. Zhao, H. Liu, and Y. Fu, "Incomplete multi-modal visual data grouping.," in *IJCAI*, pp. 2392–2398, 2016.

[114] X. Gu, J. Sun, and Z. Xu, "Spherical space domain adaptation with robust pseudo-label loss," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 9101–9110, 2020.

[115] K. Zhou, Y. Yang, T. Hospedales, and T. Xiang, "Learning to generate novel domains for domain generalization," in *European Conference on Computer Vision*, pp. 561–578, Springer, 2020.

[116] J. Dong, Y. Cong, G. Sun, Y. Liu, and X. Xu, "Cscl: Critical semantic-consistent learning for unsupervised domain adaptation," in *European Conference on Computer Vision*, pp. 745–762, Springer, 2020.

[117] R. Xu, P. Liu, L. Wang, C. Chen, and J. Wang, "Reliable weighted optimal transport for unsupervised domain adaptation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 4394–4403, 2020.

[118] Y. Chen, S. Song, S. Li, and C. Wu, "A graph embedding framework for maximum mean discrepancy-based domain adaptation algorithms," *IEEE Transactions on Image Processing*, vol. 29, pp. 199–213, 2019.

[119] A. Kumagai and T. Iwata, "Unsupervised domain adaptation by matching distributions based on the maximum mean discrepancy via unilateral transformations," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 33, pp. 4106–4113, 2019.

[120] Z. Pei, Z. Cao, M. Long, and J. Wang, "Multi-adversarial domain adaptation," in *Thirty-Second AAAI Conference on Artificial Intelligence*, 2018.

[121] X. Chen, S. Wang, M. Long, and J. Wang, "Transferability vs. discriminability: Batch spectral penalization for adversarial domain adaptation," in *International Conference on Machine Learning*, pp. 1081–1090, 2019.

[122] H.-W. Yeh, B. Yang, P. C. Yuen, and T. Harada, “Sofa: Source-data-free feature alignment for unsupervised domain adaptation,” in *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pp. 474–483, 2021.

[123] Z. Xu, S. Huang, Y. Zhang, and D. Tao, “Webly-supervised fine-grained visual categorization via deep domain adaptation,” *IEEE transactions on pattern analysis and machine intelligence*, vol. 40, no. 5, pp. 1100–1113, 2018.

[124] H. Liu, M. Shao, Z. Ding, and Y. Fu, “Structure-preserved unsupervised domain adaptation,” *IEEE Transactions on Knowledge and Data Engineering*, 2018.

[125] T. Chen, S. Kornblith, M. Norouzi, and G. Hinton, “A simple framework for contrastive learning of visual representations,” in *International conference on machine learning*, pp. 1597–1607, PMLR, 2020.

[126] M. Zhang, J. Zhang, Z. Lu, T. Xiang, M. Ding, and S. Huang, “Iept: Instance-level and episode-level pre-text tasks for few-shot learning,”

[127] S. Cui, S. Wang, J. Zhuo, L. Li, Q. Huang, and Q. Tian, “Towards discriminability and diversity: Batch nuclear-norm maximization under label insufficient situations,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 3941–3950, 2020.

[128] Y. Jin, X. Wang, M. Long, and J. Wang, “Minimum class confusion for versatile domain adaptation,” in *European Conference on Computer Vision*, pp. 464–480, Springer, 2020.

[129] Y. Kim, S. Hong, D. Cho, H. Park, and P. Panda, “Domain adaptation without source data,” *arXiv preprint arXiv:2007.01524*, 2020.

[130] C. Xu, Z. Guan, W. Zhao, H. Wu, Y. Niu, and B. Ling, “Adversarial incomplete multi-view clustering,” in *Proceedings of the 28th International Joint Conference on Artificial Intelligence*, pp. 3933–3939, AAAI Press, 2019.

[131] H. Chi, H. Xia, L. Zhang, C. Zhang, and X. Tang, “Competitive and collaborative representation for classification,” *Pattern Recognition Letters*, vol. 132, pp. 46–55, 2020.

[132] T. Yuan, F. Wan, M. Fu, J. Liu, S. Xu, X. Ji, and Q. Ye, “Multiple instance active learning for object detection,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 5330–5339, 2021.

[133] A. Kirillov, Y. Wu, K. He, and R. Girshick, “Pointrend: Image segmentation as rendering,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 9799–9808, 2020.

[134] I. Anokhin, K. Demochkin, T. Khakhulin, G. Sterkin, V. Lempitsky, and D. Korzhenkov, “Image generators with conditionally-independent pixel synthesis,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 14278–14287, 2021.

[135] Q. Wang and T. Breckon, "Unsupervised domain adaptation via structured prediction based selective pseudo-labeling," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 34, pp. 6243–6250, 2020.

[136] S. Chen, M. Harandi, X. Jin, and X. Yang, "Domain adaptation by joint distribution invariant projections," *IEEE Transactions on Image Processing*, vol. 29, pp. 8264–8277, 2020.

[137] H. Wu, H. Zhu, Y. Yan, J. Wu, Y. Zhang, and M. K. Ng, "Heterogeneous domain adaptation by information capturing and distribution matching," *IEEE Transactions on Image Processing*, vol. 30, pp. 6364–6376, 2021.

[138] K. Zhou, Y. Yang, Y. Qiao, and T. Xiang, "Domain adaptive ensemble learning," *IEEE Transactions on Image Processing*, vol. 30, pp. 8008–8018, 2021.

[139] Z. Huang, H. Wang, E. P. Xing, and D. Huang, "Self-challenging improves cross-domain generalization," *arXiv preprint arXiv:2007.02454*, vol. 2, 2020.

[140] S. Choi, S. Jung, H. Yun, J. T. Kim, S. Kim, and J. Choo, "Robustnet: Improving domain generalization in urban-scene segmentation via instance selective whitening," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 11580–11590, 2021.

[141] R. Volpi, H. Namkoong, O. Sener, J. Duchi, V. Murino, and S. Savarese, "Generalizing to unseen domains via adversarial data augmentation," *arXiv preprint arXiv:1805.12018*, 2018.

[142] Q. Dou, D. Coelho de Castro, K. Kamnitsas, and B. Glocker, "Domain generalization via model-agnostic learning of semantic features," *Advances in Neural Information Processing Systems*, vol. 32, pp. 6450–6461, 2019.

[143] Y. Yang and Z. Xu, "Rethinking the value of labels for improving class-imbalanced learning," *Advances in neural information processing systems*, vol. 33, pp. 19290–19301, 2020.

[144] J. Ren, C. Yu, X. Ma, H. Zhao, S. Yi, *et al.*, "Balanced meta-softmax for long-tailed visual recognition," *Advances in neural information processing systems*, vol. 33, pp. 4175–4186, 2020.

[145] P. Chattopadhyay, Y. Balaji, and J. Hoffman, "Learning to balance specificity and invariance for in and out of domain generalization," *Proceedings of the European Conference on Computer Vision*, 2020.

[146] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," *Communications of the ACM*, vol. 60, no. 6, pp. 84–90, 2017.

[147] Z.-S. Liu, V. Kalogeiton, and M.-P. Cani, "Multiple style transfer via variational autoencoder," in *2021 IEEE International Conference on Image Processing (ICIP)*, pp. 2413–2417, IEEE, 2021.

[148] L. V. Kantorovich, "On the translocation of masses," *Journal of Mathematical Sciences*, vol. 133, no. 4, pp. 1381–1382, 2006.

[149] I. Tolstikhin, O. Bousquet, S. Gelly, and B. Schoelkopf, “Wasserstein auto-encoders,” *arXiv preprint arXiv:1711.01558*, 2017.

[150] C. Fang, Y. Xu, and D. N. Rockmore, “Unbiased metric learning: On the utilization of multiple datasets and web images for softening bias,” in *Proceedings of the IEEE International Conference on Computer Vision*, pp. 1657–1664, IEEE, 2013.

[151] Y. Balaji, S. Sankaranarayanan, and R. Chellappa, “Metareg: Towards domain generalization using meta-regularization,” in *Advances in Neural Information Processing Systems*, pp. 998–1008, 2018.

[152] H. Zhang, M. Cisse, Y. N. Dauphin, and D. Lopez-Paz, “mixup: Beyond empirical risk minimization,” *arXiv preprint arXiv:1710.09412*, 2017.

[153] W. Lu, J. Wang, H. Li, Y. Chen, and X. Xie, “Domain-invariant feature exploration for domain generalization,” *Transactions on Machine Learning Research*, 2022.

[154] D. Krueger, E. Caballero, J.-H. Jacobsen, A. Zhang, J. Binas, D. Zhang, R. Le Priol, and A. Courville, “Out-of-distribution generalization via risk extrapolation (rex),” in *International Conference on Machine Learning*, pp. 5815–5826, PMLR, 2021.

[155] S. Seo, Y. Suh, D. Kim, J. Han, and B. Han, “Learning to optimize domain specific normalization for domain generalization,” *arXiv preprint arXiv:1907.04275*, vol. 3, no. 6, p. 7, 2019.

[156] K.-C. Peng, Z. Wu, and J. Ernst, “Zero-shot deep domain adaptation,” in *ECCV*, pp. 764–781, 2018.

[157] Z. Ding, M. Shao, and Y. Fu, “Latent low-rank transfer subspace learning for missing modality recognition,” in *Proceedings of Association for the Advancement of Artificial Intelligence*, pp. 1192–1198, 2014.

[158] K. Muandet, D. Balduzzi, and B. Schölkopf, “Domain generalization via invariant feature representation,” in *ICML*, pp. 10–18, 2013.

[159] J. Wang and J. Jiang, “Conditional coupled generative adversarial networks for zero-shot domain adaptation,” in *ICCV*, pp. 3375–3384, 2019.

[160] M. D. Zeiler, D. Krishnan, G. W. Taylor, and R. Fergus, “Deconvolutional networks,” in *CVPR*, pp. 2528–2535, IEEE, 2010.

[161] Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner, “Gradient-based learning applied to document recognition,” *Proceedings of the IEEE*, vol. 86, no. 11, pp. 2278–2324, 1998.

[162] H. Xiao, K. Rasul, and R. Vollgraf, “Fashion-mnist: a novel image dataset for benchmarking machine learning algorithms,” *arXiv preprint arXiv:1708.07747*, 2017.

[163] G. Cohen, S. Afshar, J. Tapson, and A. Van Schaik, “Emnist: Extending mnist to handwritten letters,” in *IJCNN*, pp. 2921–2926, IEEE, 2017.

[164] P. Arbelaez, M. Maire, C. Fowlkes, and J. Malik, “Contour detection and hierarchical image segmentation,” *IEEE transactions on pattern analysis and machine intelligence*, vol. 33, no. 5, pp. 898–916, 2010.

[165] C. Zhang, Y. Cai, G. Lin, and C. Shen, "Deepemd: Few-shot image classification with differentiable earth mover's distance and structured classifiers," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 12203–12213, 2020.

[166] "Augmented multi-modality fusion for generalized zero-shot sketch-based visual retrieval," *IEEE Transactions on Image Processing*, 2022.

[167] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, "You only look once: Unified, real-time object detection," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 779–788, 2016.

[168] D. Guan, J. Huang, A. Xiao, S. Lu, and Y. Cao, "Uncertainty-aware unsupervised domain adaptation in object detection," *IEEE Transactions on Multimedia*, 2021.

[169] X. Lu, W. Wang, M. Danelljan, T. Zhou, J. Shen, and L. Van Gool, "Video object segmentation with episodic graph memory networks," in *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part III 16*, pp. 661–679, Springer, 2020.

[170] H. Xia, T. Jing, and Z. Ding, "Maximum structural generation discrepancy for unsupervised domain adaptation," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2022.

[171] T. Jing, H. Liu, and Z. Ding, "Towards novel target discovery through open-set domain adaptation," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 9322–9331, 2021.

[172] M. Baktashmotlagh, M. T. Harandi, B. C. Lovell, and M. Salzmann, "Unsupervised domain adaptation by domain invariant projection," in *Proceedings of the IEEE International Conference on Computer Vision*, pp. 769–776, 2013.

[173] X. Liu, Z. Guo, S. Li, F. Xing, J. You, C.-C. J. Kuo, G. El Fakhri, and J. Woo, "Adversarial unsupervised domain adaptation with conditional and label shift: Infer, align and iterate," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 10367–10376, 2021.

[174] D. Guan, J. Huang, S. Lu, and A. Xiao, "Scale variance minimization for unsupervised domain adaptation in image segmentation," *Pattern Recognition*, vol. 112, p. 107764, 2021.

[175] R. Aljundi, R. Emonet, D. Muselet, and M. Sebban, "Landmarks-based kernelized subspace alignment for unsupervised domain adaptation," in *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, pp. 56–63, 2015.

[176] S. Ioffe and C. Szegedy, "Batch normalization: Accelerating deep network training by reducing internal covariate shift," in *International conference on machine learning*, pp. 448–456, PMLR, 2015.

[177] Q. Wang, J. Cheng, Q. Gao, G. Zhao, and L. Jiao, "Deep multi-view subspace clustering with unified and discriminative learning," *IEEE Transactions on Multimedia*, 2020.

[178] D. Elbrächter, D. Perekrestenko, P. Grohs, and H. Bölcskei, “Deep neural network approximation theory,” *IEEE Transactions on Information Theory*, vol. 67, no. 5, pp. 2581–2623, 2021.

[179] B. Fernando, A. Habrard, M. Sebban, and T. Tuytelaars, “Unsupervised visual domain adaptation using subspace alignment,” in *Proceedings of IEEE International Conference on Computer Vision*, 2013.

[180] M. Wang, W. Wang, B. Li, X. Zhang, L. Lan, H. Tan, T. Liang, W. Yu, and Z. Luo, “Interbn: Channel fusion for adversarial unsupervised domain adaptation,” in *Proceedings of the 29th ACM international conference on multimedia*, pp. 3691–3700, 2021.

[181] L. Zhang, M. Yang, and X. Feng, “Sparse representation or collaborative representation: Which helps face recognition?,” in *2011 International conference on computer vision*, pp. 471–478, IEEE, 2011.

[182] K. Lai, L. Bo, X. Ren, and D. Fox, “A large-scale hierarchical multi-view rgb-d object dataset,” in *2011 IEEE international conference on robotics and automation*, pp. 1817–1824, IEEE, 2011.

[183] A. Janoch, S. Karayev, Y. Jia, J. T. Barron, M. Fritz, K. Saenko, and T. Darrell, “A category-level 3d object dataset: Putting the kinect to work,” in *Consumer depth cameras for computer vision*, pp. 141–165, Springer, 2013.

[184] G. Griffin, A. Holub, and P. Perona, “Caltech-256 object category dataset,” 2007.

[185] Z. Du, J. Li, H. Su, L. Zhu, and K. Lu, “Cross-domain gradient discrepancy minimization for unsupervised domain adaptation,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 3937–3946, 2021.

[186] J. Na, H. Jung, H. J. Chang, and W. Hwang, “Fixbi: Bridging domain spaces for unsupervised domain adaptation,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 1094–1103, 2021.

[187] Y. Li, L. Yuan, Y. Chen, P. Wang, and N. Vasconcelos, “Dynamic transfer for multi-source domain adaptation,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 10998–11007, 2021.

[188] S. Ben-David, J. Blitzer, K. Crammer, and F. Pereira, “Analysis of representations for domain adaptation,” in *Proceedings of Advances in Neural Information Processing Systems*, pp. 137–144, 2007.

Biography

Haifeng Xia was born in Shandong, China. He received the B.S. degree in Information and Computer Science from Huazhong Agricultural University, Wuhan, China, in 2016 and received the M.S. degree in Computational Mathematics from Sun Yat-sen University, Guangzhou, China, in 2019. He started his Ph.D study in the Department of Computer Science, Tulane University in January 2021 and is completing this program in May 2023. His current research interests mainly include computer vision and machine learning.