UTS UNIVERSITY OF TECHNOLOGY SYDNEY

Cross-Domain Image Classification in Complex Real-World Scenarios: With Test-Time Label or Continual Domain Shift

by Tianyi Ma

Thesis submitted in fulfilment of the requirements for the degree of

Doctor of Philosophy

under the supervision of Maoying Qiao

University of Technology Sydney
Faculty of Engineering and Information Technology

September 2025

CERTIFICATE OF ORIGINAL AUTHORSHIP

I, *Tianyi Ma* declare that this thesis is submitted in fulfilment of the requirements for the award of Doctor of Philosophy, in the *Faculty of Engineering and Information Technology* at the University of Technology Sydney.

This thesis is wholly my own work unless otherwise referenced or acknowledged. In addition, I certify that all information sources and literature used are indicated in the thesis.

This document has not been submitted for qualifications at any other academic institution.

This research was supported by an Australian Government Research Training Program (RTP) Scholarship doi.org/10.82133/C42F-K220.

SIGNATURE: Production Note: Signature removed prior to publication. 4 Sep 2025

[Tianyi Ma]

DATE: 4th September, 2025

PLACE: Sydney, Australia

Acknowledgments

I would like to express my deepest gratitude to my supervisor, Dr. Maoying Qiao, for her continuous guidance, encouragement, and patience throughout my PhD journey. Her invaluable support has been essential, and without her mentorship, I could not have reached this stage and completed my thesis.

I am also truly grateful to my housemate, Bishal Shrestha, whose kindness and uplifting spirit have supported me through many difficult moments. My heartfelt thanks extend to my fellow PhD colleagues, Rui Li and Zekai Wang, for their emotional support and companionship along the way.

I sincerely acknowledge the University of Technology Sydney and the HDR Research School for their financial and academic support. The scholarship I received not only enabled me to pursue my research but also made my life in Australia as an international student possible, especially during times when financial challenges would otherwise have been overwhelming.

I owe profound thanks to my parents, whose love and unwavering support gave me the courage to come to Australia alone during the peak of the COVID-19 pandemic. Their sacrifices and faith in me have been a constant source of strength.

Finally, I would like to extend a word of gratitude to myself. Despite the many difficulties, uncertainties, and challenges, I persevered. It is this resilience that ultimately brought me here, and for that, I am deeply thankful.

Abstract

Cross-domain image classification is proposed to address the distribution discrepancy between the source and target domains, commonly referred to as the domain gap. Although it has been proven effective in mitigating the domain gap between the source and target domains, cross-domain image classification still faces significant challenges in complex real-world scenarios. In this thesis, two specific challenges that commonly happen, namely, label shift and continual domain shift, for cross-domain image classification are investigated and addressed. For the label shift, this thesis focuses on cross-domain few-shot image classification (CDFSIC). A novel prompt-to-disentangle method is proposed to combine the benefits of domain generalisation and adaptation by disentangling source and target knowledge. For the continual domain shift, this thesis focuses on continual test-time adaptation (CoTTA). Based on the findings in CDFSIC, we design a similar strategy called the Source and Target Disentangle Transformer to explicitly disentangle source and target knowledge, thereby facilitating both the preservation of source knowledge and the extraction of target knowledge. Then, based on the observation that the recent CoTTA method is unstable in a small-batch setting, a novel task named single-sample CoTTA is proposed. A novel strategy, named effective buffer and resetting, is designed to increase adaptation stability. Moreover, we apply this method to zero-shot models, solving the label shift and continual domain shift simultaneously. Finally, we highlighted several future directions, including active CoTTA to address larger domain gaps with human calibrations and zero-shot CoTTA to tackle both label shift and continual domain shift.

List of Publications

1. ProD: Prompting-to-disentangle Domain Knowledge for Cross-domain Few-shot Image Classification. Tianyi Ma, Yifan Sun, Zongxin Yang. The IEEE/CVF Conference on Computer Vision and Pattern Recognition 2023 (CVPR 2023)

2. Disentangle Source and Target Knowledge for Continual Test-Time Adaptation. Tianyi Ma, Maoying Qiao. IEEE/CVF Winter Conference on Applications of Computer Vision 2025 (WACV 2025)

3. EBaR: Efficient Buffer and Resetting for Single-Sample Continual Test-Time Adaptation. Tianyi Ma, Maoying Qiao. The Association for Computing Machinery's (ACM) annual international conference on multimedia (ACM MM 2025)

Table of Contents

List of Figures

List of Tables

Table **Page**

CHAPTER 1

INTRODUCTION

Deep learning [55, 91] has seen astonishing progress in recent years and is now capable of solving a wide range of tasks, including visual recognition [119, 137, 144], natural language processing [27, 70, 131], and audio recognition [73, 140, 205]. In this work, we focus primarily on the image classification task within the broader field of visual recognition. Image classification [39, 64, 118] refers to the process of assigning a label or category to an image based on its visual content. Specifically, given an input image, the goal is to predict the most appropriate class from a predefined set of categories, typically by analysing patterns, textures, shapes, and other visual features present in the image. Recent studies typically involve training a deep neural network on labelled [64] or unlabelled image data [21], often requiring a large volume of training samples to achieve satisfactory performance. Once trained, the model is either directly evaluated on or further adapted to downstream test images.

Standard deep learning methods for image classification typically rely on the assumption that both training (source) and testing (target) data are distributed independently and identically (i.i.d.) [141]. Under this assumption, models trained on a source domain are expected to generalise well to unseen data drawn from the same distribution. However, in many real-world scenarios, this assumption is often violated, as the target data encountered during deployment may come from a distribution that significantly differs from the one used during training. Such distribution shifts can arise due to changes in lighting conditions, camera sensors, background clutter, or even object appearance and style, leading to substantial performance degradation when deep learning models are

directly applied to the target domain. To address this issue, cross-domain learning for image classification has emerged as a critical research direction. It aims to improve the ability to generalise or adapt to unseen target domains, leveraging knowledge from the source domain and target domain.

1.1 Cross-Domain Learning

Cross-domain learning is proposed to address the distribution shift issue. The distribution shift causes the distribution discrepancy between the source and target domains, commonly referred to as the domain gap. The domain gap can significantly hinder the models from achieving optimum performance during application. To bridge the domain gap, two main approaches are widely studied: domain generalisation [81, 173, 219] and domain adaptation [96, 174, 185]. Domain generalisation aims to learn models that can generalise well to unseen target domains without accessing any target data during training. In contrast, domain adaptation leverages target data, labelled or unlabelled, during training or testing to reduce distribution discrepancy between source and target.

However, although cross-domain learning has proven effective in mitigating the domain gap between the source and target domains, it faces significant challenges beyond the domain gap in more complex real-world scenarios. Two commonly observed challenges are label shift and continual domain shift.

Label shift refers to the case where the label distributions differ between the source and target domains. This mismatch can lead to biased predictions and significantly reduced classification accuracy. Continual domain shift, on the other hand, describes the situation when the target distribution changes continually during deployment. This continual change poses severe difficulties for traditional cross-domain learning approaches, which typically assume static target distributions. Addressing these issues requires more robust and adaptive learning strategies.

1.2 Label Shift

The label shift [5, 31, 50, 51] refers to the situation in which the label distributions are in a discrepancy between the source and target domains. For example, label shift occurs when class labels are evenly distributed in the source domain but unevenly distributed in the target domain [85, 159]. Label shift can significantly affect the performance of deep learning models, as the models may become overconfident in predicting certain

classes while underrepresenting others, leading to biased predictions and a substantial drop in classification accuracy, particularly in scenarios where the model heavily relies on the source label prior during inference.

A more challenging situation under label shift arises when the target domain contains novel classes that have never been seen in the source domain during model training. This thesis focuses on addressing this challenging case of label shift. To handle novel classes in the target domain for image classification, two main approaches are commonly studied: few-shot image classification and zero-shot image classification.

Few-shot image classification [1, 36, 163] aims to adapt a pre-trained model to novel classes using only a very limited number of labelled examples per class, typically 1, 5, or 10 samples [153]. After being fine-tuned on these few labelled instances, a few-shot model is directly evaluated on the remaining target data belonging to the novel classes. In contrast, zero-shot image classification [138] seeks to recognise novel classes without any labelled examples from those classes during training. Zero-shot learning is currently achieved by leveraging a general vision-language model capable of aligning visual inputs with semantic descriptions, such as class names or attributes, thereby enabling the model to make predictions for previously unseen categories. In this thesis, we primarily focus on the few-shot learning setting, as it offers a more practical and tractable solution in real-world scenarios where a small number of labelled samples can be collected for new categories. The zero-shot classification setting is also considered when discussing continual domain shift scenarios.

When applying few-shot image classification under domain shift conditions, the problem becomes Cross-Domain Few-Shot Image Classification (CDFSIC) [35, 194]. The objective of this task is to adapt a source pre-trained model using only a few labelled samples of novel classes drawn from a target domain whose feature distribution differs significantly from that of the source. This setting imposes two significant challenges: (1) the model must generalise effectively across domains during training, and (2) it must efficiently capture target-specific knowledge with only a limited number of samples per class. Meeting both of these requirements is nontrivial. Most recent methods have focused primarily on addressing the first challenge, improving cross-domain generalisation. In contrast, relatively few have effectively tackled the second challenge of extracting target-specific knowledge from limited data. A detailed review of recent methods in light of these two challenges is provided in the following literature review, Chapter 2.

To this end, this thesis proposes a novel method, Prompt-to-Disentangle (ProD), as detailed in Chapter 3. ProD aims to disentangle domain-general and domain-specific

knowledge by incorporating two parallel prompts within a vision transformer architecture, referred to as the Domain-General (DG) prompt and the Domain-Specific (DS) prompt. The DG prompt is trained to capture knowledge that is invariant across domains and transferable to downstream target domains, thereby addressing the first requirement of cross-domain generalisation. In contrast, the DS prompt is designed to rapidly extract target-specific features and apply them as conditioning information in an on-the-fly manner. The DS prompt allows the model to efficiently adapt to novel target domains using only a few labelled samples, satisfying the second requirement of effective target-specific knowledge extraction. Comprehensive ablation studies and empirical evaluations demonstrate the effectiveness of both components of ProD. Furthermore, the joint use of the DG and DS prompts enables ProD to achieve SOTA performance on the CDFSIC task under various settings across multiple benchmark datasets.

1.3 Continual Domain Shift

The continual domain shift [19, 100, 170, 182, 209, 211] refers to the phenomenon in which the distribution of the target domain changes continually at the test time. This continual shift in data distribution commonly occurs in real-world scenarios. For instance, in autonomous driving, the surrounding environment, such as lighting, road conditions, and traffic, evolves constantly as the vehicle moves through different scenes. Similarly, in surveillance applications, factors such as weather conditions, lighting variations, and fluctuations in human activity levels can lead to a continuously changing input distribution under CCTV monitoring. These scenarios challenge the assumption of a fixed target distribution and require models to adapt in real time to maintain robust performance.

To address the challenge of continual domain shift, recent studies have proposed the Continual Test-Time Adaptation (CoTTA) [175]. CoTTA extends the standard test-time adaptation (TTA) [100] task. Similar to TTA, CoTTA requires the model to adapt to target domain during the testing phase with unlabelled target data. The key distinction lies in the target domain: while standard TTA assumes a fixed target distribution, CoTTA considers a dynamic setting where the target domain distribution changes continually over time. As a result, models must not only adapt to the target in an unsupervised manner but also maintain performance across a long stream of test data from changing domains. In this thesis, we focus on a more challenging and realistic setting: source-free CoTTA. Under the source-free setting, the source data is unavailable when testing. This

constraint reflects practical scenarios where storage, privacy, or regulatory concerns prevent retaining source domain samples.

Source-free CoTTA (referred to as CoTTA in the following sections) has two key challenges. The first is how to effectively preserve source domain knowledge over an extended period without access to the source data. When adapting to continually changing target domains, the model must retain the domain-shared source knowledge embedded in the pre-trained source model. If this knowledge is discarded during the adaptation process, a phenomenon known as catastrophic forgetting occurs. Catastrophic forgetting is a well-known challenge in the field of continual learning [33, 183, 183], where a model gradually forgets previously learned tasks when trained sequentially on new ones. While both CoTTA and continual learning suffer from catastrophic forgetting, there is a fundamental difference between the two. In continual learning, the model is exposed to a sequence of explicitly defined tasks, often accompanied by new labels and objectives. In contrast, CoTTA does not introduce new tasks or labels.

The second challenge is to extract target-specific knowledge from the continually evolving target domain effectively. As the target distribution shifts over time, the model must quickly adapt to these changes to maintain high performance. This requirement becomes even more demanding in the absence of target labels, as the model must rely entirely on the incoming unlabelled data to guide its adaptation. Moreover, the target samples encountered during test time can be highly noisy or distributed far outside the distribution seen during pre-training. Such out-of-distribution or low-quality samples make it difficult for the model to learn meaningful and generalisable representations. Consequently, extracting high-quality and domain-specific knowledge from unlabelled, non-stationary, and potentially corrupted input poses a significant obstacle in the source-free CoTTA setting.

Recent works on CoTTA are reviewed in Chapter 2. Most existing methods primarily focus on addressing only one of the two core challenges. Even when some approaches attempt to tackle both challenges jointly, they often solve them without explicitly disentangling the two. This thesis argues that a more effective strategy is to address the two challenges in a disentangled manner, allowing each component of the model to specialise in a distinct aspect of the adaptation process. In this way, both challenges can be addressed more effectively and explicitly.

To this end, this thesis proposes a novel method, named Source and Target Disentangle Transformer (SoTa-DiT), which explicitly disentangles source and target knowledge to address the two key challenges. The details of SoTa-DiT are provided in Chapter 4.

Inspired by the success of the ProD method, SoTa-DiT adopts a dual-prompt mechanism within a vision transformer architecture. Specifically, one set of prompts, referred to as the source prompts, is designed to extract and preserve source domain knowledge. The source prompt is supervised by a carefully designed source preservation loss, which ensures that critical domain-shared information is retained during continual adaptation. In parallel, another set of prompts, called the target prompts, focuses on extracting high-quality, target-related knowledge from the unlabelled target data of continually changing domains. The target prompt is supervised by a group of target extraction losses, which encourage the model to learn discriminative and robust features for the current target distribution. The final classification prediction is obtained by combining the outputs from both the source and target prompts, allowing SoTa-DiT to leverage both source and target knowledge. SoTa-DiT achieves SOTA performance across multiple benchmark datasets and transformer backbones. Furthermore, comprehensive ablation studies and empirical analyses are conducted to validate the effectiveness of each component of SoTa-DiT. These experiments demonstrate: (1) the contribution of each prompt to overall performance, and (2) the disentangle effect of SoTa-DiT. The results strongly support our central hypothesis that addressing the two challenges of source-free CoTTA in a disentangled manner leads to more robust and effective solution.

This thesis further investigates beyond the standard CoTTA setting. We observe that many recent CoTTA methods exhibit instability and often collapse when the testing batch size is small. This issue arises because noisy or out-of-distribution samples can introduce significant perturbations to the model during adaptation, especially when the samples are processed individually or in small batches. To address this issue, this thesis introduces a novel task, termed Single-Sample CoTTA (S-CoTTA), which requires the model to perform CoTTA under the extreme case where the test batch size is equal to one. Under the S-CoTTA setting, standard CoTTA models frequently collapse due to the destabilising effect of noisy individual samples, while existing single-sample test-time adaptation methods [149] tend to lack long-term stability and fail to maintain performance over time. Therefore, solving the S-CoTTA task presents two essential requirements: (1) the model must remain robust and stable under small-batch or single-sample settings, and (2) it must continuously adapt to the changing target domains without catastrophic forgetting over a long time step.

To this end, this thesis further proposes a novel method, named Efficient Buffer and Resetting (EBaR). EBaR is a general method that is compatible with both convolutional neural networks (CNNs) and transformer-based architectures. The details of EBaR are

presented in Chapter 5. EBaR consists of two key components: an efficient buffer and an elastic resetting mechanism. The efficient buffer module comprises two types of buffers: a high-uncertainty buffer, which is larger and stores samples with high uncertainty, and a low-uncertainty buffer, which is smaller and stores samples with low uncertainty. Once the buffers reach capacity, samples are popped and different loss functions are applied selectively to update the model. This design offers two key benefits: (1) samples are accumulated and aggregated over time, thereby mitigating the perturbation effect of individual noisy samples; and (2) uncertainty-aware updating stabilises the adaptation process by giving less attention to highly uncertain and noisy samples. The second component, elastic resetting, helps maintain stability and prevents catastrophic forgetting by selectively resetting the model parameters toward their original values. For each parameter, the elastic resetting unit tracks its value change over time. If the change of a parameter exceeds a predefined threshold, it is reset using a dynamic weighted reset strategy. This strategy resets parameters that are more sensitive to perturbations closer to their original values, while leaving others more flexible. As a result, the model is able to preserve essential source knowledge while still retaining its ability to adapt to new target domains. Experimental results demonstrate that EBaR achieves SOTA performance across multiple benchmark datasets with various backbones. Comprehensive ablation studies and observation-based experiments are conducted to validate the effectiveness of each individual component within EBaR. These analyses confirm that both the efficient buffer and the elastic resetting mechanisms contribute significantly to overall performance. Furthermore, the results prove that EBaR effectively addresses the two key challenges posed by the S-CoTTA setting: maintaining stability under single-sample conditions and preserving source knowledge during long-term adaptation. In addition, zero-shot models are also evaluated in the S-CoTTA setting using EBaR and achieve significantly improved classification accuracy. These results indicate that EBaR generalises well to zero-shot models. In this way, the proposed framework addresses both the label shift problem and the continual domain shift problem in a unified manner.

1.4 Jointly Solving Label and Continual Domain Shift

Jointly addressing label shift and continual domain shift is of substantial practical significance for cross-domain learning. In complex real-world scenarios, deep learning models frequently encounter both previously unseen classes and continually evolving target domains. For example, in CCTV-based visual analysis systems, models trained

on source-domain images or videos may be deployed in environments where visual data undergo continual domain shifts due to changes in lighting conditions, weather, camera viewpoints, and human density. Similar challenges arise in other applications, such as disease detection, where novel disease categories may emerge with variations in environmental conditions. In autonomous driving, unseen objects may appear under dynamically changing weather and traffic patterns. Other safety-critical perception tasks face comparable issues. These scenarios highlight that label shift and continual domain shift often occur together in practice. Addressing both challenges together is critical for robust cross-domain learning. This is especially important when large-scale open-world models are impractical due to computational, deployment, or resource constraints.

In this thesis, the two challenges are first investigated separately. Novel methods are designed to address each problem. In the final proposed approach, which primarily targets continual domain shift, a zero-shot model is incorporated to jointly address label shift and continual domain shift. Experimental results show that the proposed method improves zero-shot image classification accuracy under both shifts. This provides a unified solution that simultaneously mitigates their effects.

1.5 Thesis Structure

The structure of this thesis is illustrated in Figure 1.1. This thesis addresses two key problems commonly encountered in real-world cross-domain image classification scenarios: label shift and continual domain shift. To address label shift, the thesis focuses on the CDFSIC task. A brief review is first presented to summarise recent progress and identify research gaps in the field (Chapter 2). Based on the insights, the first proposed method, ProD: Prompt-to-Disentangle, is introduced to effectively tackle the core challenges of CDFSIC (Chapter 3). For continual domain shift, the thesis focuses on the CoTTA task. A comprehensive survey is conducted to summarise recent progress in this area (Chapter 2). Inspired by the gap we find in the survey and the effectiveness of ProD, the second proposed method, SoTa-DiT: Source and Target Disentangle Transformer, is introduced to address the CoTTA task by explicitly disentangling source and target knowledge (Chapter 4). Following this, to overcome the instability observed in recent CoTTA methods under single-sample conditions, where the test batch size is set to one, the third proposed method, EBaR: Efficient Buffer and Resetting, is introduced (Chapter 5). Moreover, EBaR is applied to zero-shot models to jointly address both the label shift and the continual domain shift. Finally, future research directions are discussed

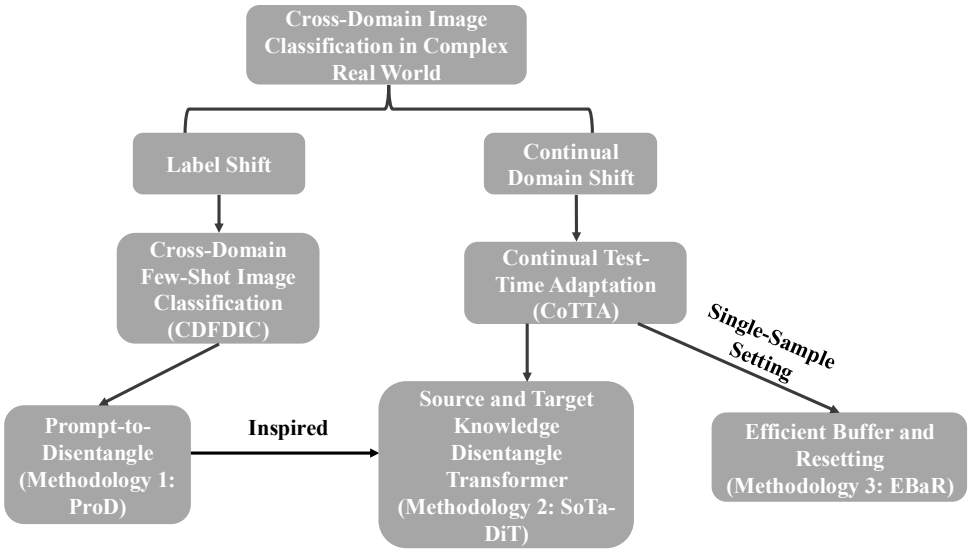


Figure 1.1: The overall thesis structure. Two specific problems are addressed for cross-domain image classification: label shift and continual domain shift. To tackle label shift, this thesis focuses on the cross-domain few-shot learning task, accompanied by a brief review and the proposed method, ProD (Contribution One). For continual domain shift, this thesis focuses on the CoTTA task, supported by a survey, the proposed methods SoTa-DiT (Contribution Two), which is inspired by ProD, and EBaR, proposed for single-sample CoTTA (Contribution Three). Further, EBaR is applied to zero-shot models to address both label shift and continual domain shift jointly.

(Chapter 6), and the conclusions of the thesis are drawn (Chapter 7).

1.6 Contributions

The three key contributions of this thesis can be summarised as follows:

- First, a novel method named ProD: Prompt-to-Disentangle, is proposed to address the CDFSIC task. ProD is the first approach that leverages a prompt-tuning mechanism to disentangle domain-general and domain-specific knowledge within the CDFSIC setting explicitly. The prompt designed to capture domain-general knowledge is responsible for producing transferable, generalisable features. In contrast, the prompt aimed at extracting domain-specific knowledge produces features related to the target domain. The domain-specific features are used as

conditional information to facilitate rapid adaptation. Using two prompts together combines the strengths of both domain generalisation and domain adaptation. As a result, ProD effectively mitigates the label shift caused by novel classes in cross-domain learning scenarios and achieves SOTA performance on multiple benchmark datasets.

- Second, a novel method named SoTa-DiT: Source and Target Disentangle Transformer, is proposed to address the CoTTA task. SoTa-DiT is the first approach to leverage prompt tuning within a transformer architecture to disentangle source and target knowledge for CoTTA explicitly. The source prompts, which focus on preserving source domain-shared knowledge, are supervised by a set of carefully designed source knowledge preservation losses. In parallel, the target prompts, responsible for extracting domain-specific knowledge from unlabelled target data, are guided by a group of target knowledge extraction losses. We demonstrate that explicitly disentangling source and target knowledge facilitates both the preservation of critical source knowledge and the effective extraction of target-specific knowledge. By combining these two benefits, SoTa-DiT addresses the challenges of continual domain shift and achieves SOTA performance across multiple benchmark datasets.

- Third, a novel method named EBaR: Efficient Buffer and Resetting, is proposed to address a newly defined task called Single-Sample CoTTA (S-CoTTA). S-CoTTA extends the CoTTA setting by requiring the model to adapt under the extreme case where the test-time batch size is one. EBaR consists of two key components: (1) an efficient buffer that separates incoming samples based on their prediction uncertainty and applies different loss functions accordingly to stabilise the adaptation under single-sample conditions, and (2) an elastic resetting unit that prevents forgetting by resetting model parameters toward their original pre-trained values. The elastic resetting is guided by both the accumulated parameter changes and their sensitivity to perturbations, enabling the model to retain source knowledge while maintaining adaptability. By combining these two components, EBaR achieves SOTA performance across multiple benchmark datasets and backbones. Furthermore, EBaR is applied to zero-shot models to address the label shift and continual domain shift problems jointly.

CHAPTER 2

LITERATURE REVIEW

In this chapter, recent related works on the two core tasks addressed in this thesis: Cross-Domain Few-Shot Image Classification and Continual Test-Time Adaptation, are reviewed, categorised, and briefly summarised. In addition, research gaps are discussed, which motivates the methodologies in the following chapters.

2.1 Cross-Domain Few-Shot Image Classification

2.1.1 Introduction

Few-shot learning (FSL) [110, 134, 153, 181] is proposed as a promising solution to the data scarcity problem in machine learning, particularly for image classification. Unlike traditional supervised learning approaches that require extensive labelled datasets [34, 88, 88], FSL focuses on training models capable of generalising to new labels using limited labelled examples. This paradigm is especially relevant in practical scenarios where data collection and annotation are expensive, time-consuming, or constrained by domain-specific limitations.

Although conventional FSL approaches assume that the source and target data share the same distribution [36, 145, 163], this assumption rarely holds in real-world applications. In many cases, models must adapt to tasks where the new data is from a domain that is significantly different from the source domain in the training phase. In this setting, FSL evolves to a more complicated Cross-Domain Few-Shot Learning

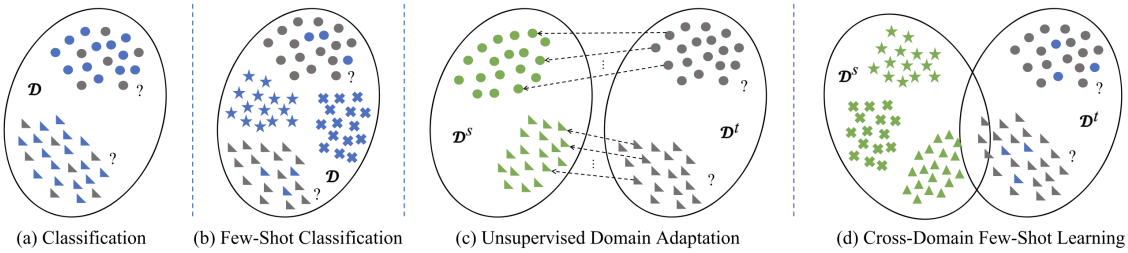


Figure 2.1: The concept of cross-domain few-shot image classification, depicted in [194]. The different shapes represent different classes. $\mathscr{D}$ means domain. $\mathscr{D}^S$ and $\mathscr{D}^T$ represent source domain and target domain, respectively. As shown in (d), cross-domain few-shot image classification jointly considers the situation in (b) and (c), introducing both domain shift and label shift introduced by novel classes, with only limited labelled samples per novel class.

(CDFSL) [35, 62, 146, 194]. CDFSL aims not only to learn novel class knowledge at the test time with limited labelled data, but to bridge the distributional gap between the source and target domains. To achieve these, the CDFSL model needs to effectively facilitate the source to increase generalisation ability while efficiently using the limited labelled target to learn target-specific knowledge for the novel classes.

In this thesis, we solve CDFSL for the image classification task, referred to as Cross-Domain Few-Shot Image Classification (CDFSIC) [35, 194]. The objective of CDFSIC is twofold: (1) to train a generalizable classification model that can perform robustly across a wide range of unseen target domains, and (2) to enable rapid adaptation to these novel domains with previously unseen class labels, using only a few labelled samples per novel class. This dual requirement makes CDFSIC particularly challenging, as it demands both strong cross-domain generalization and efficient target-specific adaptation under severe data constraints.

2.1.2 Problem Definition

Cross-Domain Few-Shot Image Classification (CDFSIC) extends the standard few-shot image classification to more realistic scenarios where the source and target domains differ. The concept of CDFSIC is illustrated in Figure 2.1. In this setting, the goal is to leverage a model trained on a labelled source domain to rapidly adapt to a target domain with novel class labels, using only limited labelled examples per novel class. The following formal definition outlines the key assumptions and objectives of the CDFSIC task.

Definition 1. ***Definition of Cross-Domain Few-Shot Image Classification (CDFSIC).***

Let $\mathscr{D}_s = \{(x_i, y_i)\}_{i=1}^{N_s}$ *be a* source domain *drawn from a joint distribution* $p_s(x,y)$*, where each* x_i *is a labelled sample from one of* C_s *base source classes. Let* $\mathscr{D}_t = \{(x_j^{(k)}, y_j^{(k)})\}_{j=1}^{K}$ *for* $k = 1,\ldots,C_n$ *denote a* target domain *drawn from a different distribution* $p_t(x,y)$*, consisting of only K labelled examples per class from a set of* C_n *novel classes, where* $p_s(x,y) \neq p_t(x,y)$ *and the label sets are disjoint:* $\mathscr{Y}_s \cap \mathscr{Y}_t = \emptyset$*.*

A learning problem is called Cross-Domain Few-Shot Image Classification (CDFSIC) *if:*

1. ***Domain shift:*** *The source and target domains follow different distributions, i.e.,* $p_s(x,y) \neq p_t(x,y)$*, indicating a domain gap.*
2. ***Label shift:*** *The target domain contains novel classes unseen in the source domain, i.e.,* $\mathscr{Y}_t \cap \mathscr{Y}_s = \emptyset$*.*
3. ***Few-shot constraint:*** *The target domain provides only a few labelled examples per novel class (commonly 1, 5, or 10), i.e.,* $K \ll N_s$*.*
4. ***Objective:*** *Given a model pre-trained on* $\mathscr{D}_s$*, adapt it using the small support set* $\mathscr{D}_t$*, and classify unseen query samples from the novel classes with high accuracy.*

2.1.3 Generalization-Based Methods

2.1.3.1 Data Augmentation

Data augmentation is commonly used during the training phase of Cross-Domain Few-Shot Image Classification (CDFSIC) to enhance the generalisation ability. By augmenting the training data without altering the semantic content, the diversity of the training set is effectively increased. This expanded diversity helps the model become more robust to typical domain shifts, such as corruption, style variation, and visual noise. For example, [154] improves generalisation by transferring the visual style of semantically similar categories to each class during training. [193] incorporates unlabelled data from auxiliary source domains by generating synthetic images that retain the content of the primary source domain but adopt the styles of the auxiliary domains. Other works, such as [54, 213], focus on enhancing diversity by extracting high-frequency components from source images and using them for augmentation. Similarly, [111] augments training data by modifying images in the frequency domain and then transforming them back into the

spatial domain for use as synthetic inputs. These augmentation strategies collectively contribute to increasing the variability of the training data and significantly improve the ability of the model to generalise across potential target domains.

2.1.3.2 Meta-Learning

Meta-learning methods adopt metric learning and episodic tasks to mimic few-shot scenarios during training and refine the optimisation process [23, 44, 150, 158, 168, 169]. Within this framework, models are trained on synthetic tasks sampled from source domains so that they can effectively generalise to novel classes and unseen domains with only a small support set. Under the CDFSIC setting, CosML [135] trains separate meta-learners for each source domain and averages their parameters during few-shot testing to improve generalisation to new domains. [162] incorporates adversarial learning into a model-agnostic meta-learning architecture to enhance domain robustness. [71] proposes a dual-adjustment mode meta-learning framework that refines class prototypes and metric selection to reduce generalisation error. [94] combines advances in transfer learning and meta-learning to create XDNet, an architecture that supports cross-domain adaptation using non-parametric classifiers while minimising computational cost. [66] introduces a parametric prototype generation mechanism based on concatenated support set features and enforces prototype-based regularisation to improve query set performance. [218] decomposes each query image into high- and low-frequency components and feeds them concurrently into the embedding network, reinforcing category prediction consistency through meta-learning. These meta-learning approaches primarily enhance the generalisation capability of models in CDFSIC. Beyond that, XDNet [94] not only improves transferability through meta-learning but also incorporates an adaptation mechanism during test time. In the proposed method, ProD, ideas from meta-learning are incorporated and reformulated as a domain-specific prompt that provides the model with informative domain-specific knowledge. By encoding meta-learning principles into a prompt-based representation, ProD enables the model to leverage prior adaptation experience while operating directly on target-domain data.

2.1.3.3 Multi-source Training

Multi-source training has become a fundamental and widely adopted strategy in the field of domain generalization [47, 156, 173, 216, 219]. The core idea behind this approach is to train models using data drawn from multiple labelled source domains. Multi-source training increases the diversity of the training domain, allowing the model to

learn more robust and domain-invariant representations. It ultimately improves the ability of a trained model to generalise and maintain performance when facing unseen domains at test time. Multiple-source training has been adopted by recent CDFSIC methods, where models must handle the domain gap with very limited labelled target data available. Liang et al. [102] propose a multi-source training approach that employs an attention mechanism to achieve effective fusion of domain-specific knowledge for CDFSIC. Similarly, Hu et al. [72] introduce a domain-switching training method that leverages multiple source domains to train a generalised teacher model capable of guiding few-shot adaptation. In another line of work, Ye et al. [200] develop an adaptive strategy that assigns different weights to conditional adversarial losses across classes, depending on the level of inter-domain discrepancy observed during training. Moreover, Liu et al. [104] propose a multi-head architecture in which each head learns representations specific to an individual source domain, while a shared backbone promotes generalisation across domains. These methods collectively demonstrate that multi-source training serves as a practical and versatile paradigm for improving the generalisation ability of models in CDFSIC tasks. The idea of multi-source learning is adopted in our proposed work, ProD, to improve the model's generalization ability.

2.1.4 Adaptation-Based Methods

2.1.4.1 Feature Transformation

In CDFSIC, feature transformation refers to the process of modifying the learned feature space from the source domain to align it with the target domain distribution better. Rather than learning entirely new representations, these methods focus on transforming existing source features to make them usable for downstream tasks. For example, [212] introduces an adaptive transformation mechanism that evaluates domain discrepancy in a task-adaptive manner, allowing the model to perform context-sensitive feature alignment. In another approach, [203] integrates both comparison-based and induction-based strategies to generate inductive meta-points that abstract and transform the original features into a more generalisable form. [20] explicitly implements feature transformation through a dedicated module that applies scale and shift operations to recalibrate the features for cross-domain alignment. [108] utilises a high-dimensional geometric algebra framework to map source features into a transformed space that is more compatible with the target distribution. Through such transformation-based strategies, the representations learned from source domains can be effectively adapted

to target domains, even when only a few labelled examples are available. These methods offer flexible and powerful strategies to bridge domain gaps in CDFSIC, eliminating the need for abundant labels. In ProD, the proposed work of this thesis, the feature transformation is done implicitly in the model layer with the provided target-domain sample features in the domain-specific prompt.

2.1.4.2 Feature Alignment

In CDFSIC, feature alignment also plays a crucial role in reducing the domain gap between the source and target domains. The objective is to adjust these feature distributions so that they become more consistent, effectively narrowing the domain gap. This alignment process facilitates the transfer of knowledge learned from the source domain to the target domain, even when only a few labelled examples are available. For instance, [59] introduces a bridge domain into which both source and target features are projected, allowing for the alignment of features through a shared intermediate representation. [195] and [199] adopt normalisation calibration strategies to align the statistical properties of the source and target domains, making feature distributions more comparable. [22] enhances alignment by simultaneously learning prototypical compact representations and enforcing bidirectional consistency between domains, resulting in more robust cross-domain representations. Similarly, [132] proposes a bidirectional cross-attention mechanism to extract transferable features that bridge the domain gap. [192] takes a different route by employing information maximisation along with distance-aware contrastive learning to implicitly align the characteristics of the source and the target without requiring explicit supervision. Collectively, these methods demonstrate the effectiveness of aligning feature distributions between domains, even under the few-shot constraint, by leveraging either explicit alignment strategies or implicitly guided learning signals. As a result, they significantly enhance the ability of the models to adapt to target domains in CDFSIC settings. In ProD, the feature alignment is also done implicitly in the model layer with the domain-specific prompt.

2.1.5 Research Gaps

Recent methods for CDFSIC focus more on the domain generalisation perspective. A large portion of existing work emphasises enhancing the generalisation ability of models during the training phase through strategies such as data augmentation, meta-learning, and multi-source training. These techniques aim to expose models to diverse conditions and

encourage the learning of domain-invariant representations. However, a growing yet still limited body of work explores the use of domain adaptation techniques, such as feature transformation and feature alignment. Many of these adaptation-based approaches tend to be task-specific and are often not transferable to general-purpose CDFSIC. Moreover, only a few studies attempt to address the CDFSIC problem by jointly considering both domain generalisation and domain adaptation within a unified framework.

To address these limitations, a novel method is proposed in Chapter 3, termed ProD, which introduces a dual-prompt architecture specifically designed for the CDFSIC task. ProD explicitly disentangles domain-general and domain-specific knowledge through a prompt-tuning mechanism that, to the best of our knowledge, has not been previously applied in this setting. The domain-general prompts are trained to capture knowledge that generalises across domains, aligning with domain generalisation objectives. In contrast, the domain-specific prompts are optimised to adapt to target-specific features, aligning with domain adaptation principles. By integrating both components within a single model, ProD benefits from the strengths of both and achieves SOTA performance across multiple CDFSIC benchmarks.

2.2 Continual Test-Time Adaptation

2.2.1 Introduction

Standard machine learning methods for image classification commonly rely on the assumption that both the training (source) and testing (target) samples are independently and identically distributed (i.i.d.) [141]. However, this assumption often fails in real-world application scenarios, where the testing data is drawn from a distribution that differs from the training data. For instance, in autonomous driving, lighting, weather, and road conditions may change over time [157]; or in medical imaging, variations may arise due to differences in cameras, scanning equipment, or hospital-specific protocols [2, 58, 190]. Such distribution shifts significantly hinder the ability of machine learning algorithms to maintain optimal performance. To address these challenges, researchers have proposed various approaches to enhance generalisation and adaptation capabilities. Domain Generalisation (DG) [9, 173, 219] aims to train models that can generalise well to target domains not seen during training. In contrast, Domain Adaptation (DA) [30, 43, 156, 174, 185] focuses on adapting a pre-trained model to a new target domain using test data with limited or no labels. The DA and DG methods can effectively mitigate the domain

gap between the source and fixed targets.

However, in real-world scenarios, the target domains often do not follow the fixed-distribution assumption. For example, in autonomous driving, lighting, weather, and environmental conditions change continually throughout the day. Similarly, in medical imaging, different doctors with varying habits may take scans at different times, introducing variability. We summarise these phenomena as continual domain shift (CDS). CDS in target domains pose a more profound challenge to the stability and performance of machine learning models over an extended period. Specifically, it poses two key challenges:

- Effectively prevent the catastrophic forgetting by preserving the source knowledge inside the model over a long time.

- Effectively transfer the existing knowledge to continually shifting target domains and extract high-quality target domain knowledge using unlabelled target data with potential noise.

To effectively address the challenges posed by CDS, recent research proposes a novel task called Continual Test-time Adaptation (CoTTA), which is gaining increasing popularity and attention. In CoTTA, the model typically does not have access to or only has limited access to labelled source data. Meanwhile, the model needs to adapt itself to a target domain that is suffering from CDS over time. More importantly, the target domain data is unlabelled, making the task more challenging. Still, due to its broad application potential and values, the CoTTA task has become one of the hottest topics under domain adaptation.

Here, as shown in Figure 2.2, we categorise CoTTA methods into two main types: the first type focuses on preserving the source knowledge and preventing knowledge from being forgotten, which is referred to as the anti-forgetting method. Under the anti-forgetting method, there are several sub-classes: normalisation calibration, regularisation, ensemble and distillation, source resetting, and adapter tuning. The second type focuses on effectively transferring knowledge to a continually shifting target without target labels, known as the target-adaptation method. Under the target-adaptation method, there are also several sub-classes: data augmentation, pseudo-label learning, optimisation objective design, and gradient adjustment.

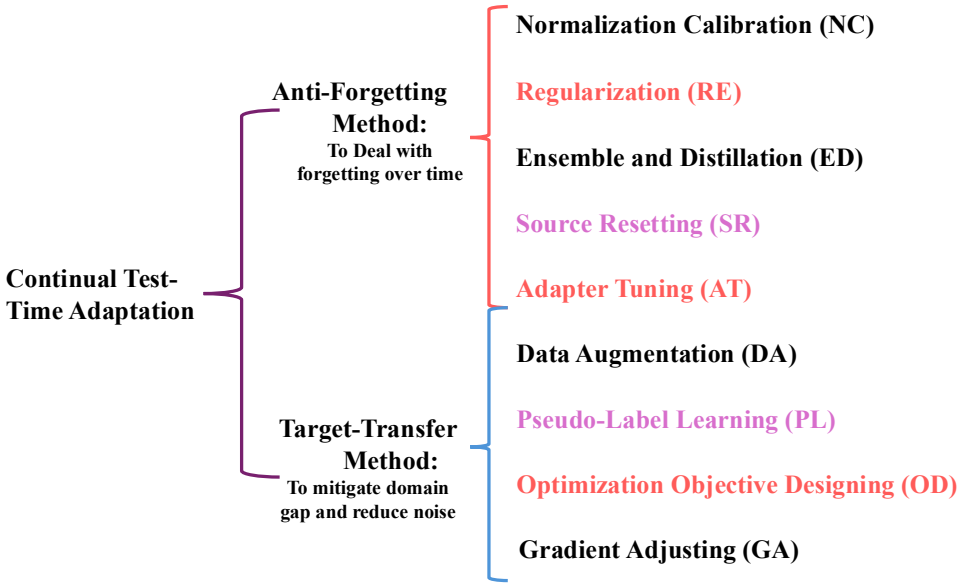


Figure 2.2: The taxonomy of continual test-time adaptation methods. The key contribution of this thesis is to consider anti-forgetting and target-adaptation jointly. For instance, Methodology 2 (SoTa-DiT) combines regularisation and adapter tuning with optimisation objective design, while Methodology 3 (EBaR) combines source resetting with pseudo-label learning.

2.2.2 Problem Definition

In the domain-adaptation setting, we often assume access to labelled source data, while the target domain is either unlabeled or only sparsely labelled. When the target distribution changes continually at deployment time, i.e., it undergoes Continual Domain Shift (CDS), and no target labels are available, the problem becomes Continual Test-Time Adaptation (CoTTA). Here, the focus is drawn on a source-free setting, where the source data is unavailable at the test time. CoTTA seeks to maintain reliable performance by adapting the model online, using only the streaming, unlabelled target data, while preserving the knowledge from the source.

Definition 2. ***Definition of Continual Test-Time Adaptation (CoTTA).***

Let $\mathscr{D}_s$ denote a source domain *with joint distribution $p_s(x,y)$ and abundant labelled data, and let $\{\mathscr{D}_t^1, \mathscr{D}_t^2, \ldots, \mathscr{D}_t^T\}$ denote a sequence of* target domains *observed at test time, each associated with an* unlabelled *joint distribution $p_t^1(x,y), \ldots, p_t^T(x,y)$. A learning problem is called* Continual Test-Time Adaptation (CoTTA) *if:*

1. ***Distribution shift:*** *$p_s(x,y) \neq p_t^t(x,y)$ for at least one t, and $p_t^t(x,y) \neq p_t^{t+1}(x,y)$ for some consecutive t, i.e., the target distribution changes continually over time (CDS).*

2. ***Label availability:*** *No ground truth labels from any target domain $\mathscr{D}_t^t$ are available during adaptation; only unlabelled target samples arrive sequentially.*

3. ***Objective:*** *Given a model trained on $\mathscr{D}_s$, continually adapt it online, using only the unlabelled target data, to minimise the prediction error on each $\mathscr{D}_t^t$ while preventing catastrophic forgetting.*

2.2.3 Anti-Forgetting Methods

2.2.3.1 Normalization Calibration

Normalisation layers in deep learning models are designed to enhance the generalisation ability by regulating the statistical properties of the data within a mini-batch. These layers help stabilise and accelerate training by normalising feature distributions. There are four primary types of normalisation layers [187]: batch normalisation (BatchNorm, BN), layer normalisation (LayerNorm, LN), group normalisation (GroupNorm, GN), and instance normalisation (InstanceNorm, IN). Batch Normalisation (BatchNorm) [76] is the most commonly used normalisation layer in CNN-based visual recognition models. It helps mitigate the risks of gradient explosion and vanishing during the training of deep neural networks.

The statistics captured by the normalisation layers are often considered closely related to domain-specific knowledge. For example, modifying normalisation layers can effectively change the style of an image [74, 80, 123]. Consequently, in the domain adaptation field, calibrating normalisation layers has emerged as an effective strategy for adapting models to new domains that exhibit different statistical properties. Notably, when only the normalisation layers are adjusted, most of the parameters remain unchanged, allowing the model to retain domain-shared knowledge learned from the source domain, which is particularly beneficial in continual test-time adaptation (CoTTA) settings. TENT [170] is the first to propose updating the statistics and affine parameters of normalisation layers for CoTTA. This approach is later extended by DUA [120] and DELTA [215]. These methods focus solely on updating the BN layers during CoTTA, without introducing additional parameters or modifying the original structure of the normalisation layers. As a result, they are widely adopted as default parameter tuning strategies in many subsequent CoTTA methods, such as EATA [128], RDUMB [139], and EATAC [160]. In the third work, EBaR, normalisation calibration is applied.

2.2.3.2 Regularization

In domain adaptation for image tasks, regularisation refers to certain constraints that guide the training process to ensure better alignment between the source domain and the target domain [5, 28, 41]. The primary objective of regularisation for CoTTA is to preserve knowledge from the source domain. Knowledge preservation is achieved by constraining the adapted parameters to remain close to those of the source model. Such constraints maintain the generalisation ability over time.

EATA [128] introduces an anti-forgetting strategy by adding a regularisation term that keeps the parameters close to the original. This constraint preserves critical weights and thus maintains source-domain performance, although an overly strong weight constraint could potentially limit adaptability to significant distribution shifts. PETAL [10] introduces a regulariser similar to EATA under a teacher-student distillation setting, pulling the student's parameters toward the source model. ECOTTA [152] takes a different approach by regularising the model indirectly using a frozen copy of the source model as a teacher. Although this method effectively preserves source knowledge and prevents drift, it still limits the ability to adapt. VCOTTA [114] introduces a variational approach to CoTTA, which regularises parameter updates based on the history of past updates, avoiding drastic updates. In the second work, SoTa-DiT, a regularisation loss is applied to ensure that the tuned prompt maintains a similar shape to the source prompt.

2.2.3.3 Ensemble and Distillation

Ensemble techniques aim to improve prediction performance by combining multiple models. The approaches of ensemble include parameter ensemble, which averages model parameters across training iterations or model instances [12, 14, 89, 90, 161], and prediction ensemble, where final predictions are obtained by voting or averaging the outputs of multiple models [3]. Distillation refers to the process of transferring knowledge from a stronger or more general model, named as the teacher, to a simpler or specialised model, named as the student [56]. Both ensemble and distillation strategies have been extensively used in semi-supervised learning [4, 14, 15, 25, 90, 130, 143, 188, 196]. They play crucial roles in enabling effective knowledge transfer, stabilising adaptation, and enhancing generalisation in settings with limited labelled data.

Inspired by advances in semi-supervised adaptation, recent CoTTA methods have also adopted ensemble and distillation techniques. Although these two strategies represent distinct approaches, they are often employed jointly in CoTTA with a shared objective:

to stabilise adaptation and prevent forgetting. Given their complementary roles and frequent integration, we discuss them together in the same section.

A representative ensemble and distillation method is COTTA [175]. In COTTA, the student model is fine-tuned through backpropagation, while the teacher model is updated using an exponential moving average (EMA) to temporally ensemble the parameter trajectory of the student. In this framework, the teacher model is updated more slowly than the student, helping retain domain-invariant source knowledge and transferring it effectively to the student. However, over a longer time, the teacher model still drifts from the source distribution, leading to forgetting. SANTA [17] introduces a source-anchoring self-distillation, using the source model as an anchoring teacher to preserve previously learned semantics, although this can limit adaptation flexibility. MJT [176] proposes a multiple teacher strategy, replacing a single EMA teacher with an ensemble of teachers trained on different source domains. DDA [49] employs a generative diffusion model to project test inputs toward the source domain and uses a self-ensemble classifier to adjust adaptation strength based on the output of the diffusion model. The strategy avoids overfitting and forgetting, but it depends on an external diffusion model and is only suitable for corruption-style shifts. In the second work, SoTa-DiT, a regularisation loss is applied to ensure that the tuned prompt maintains a similar shape to the source prompt.

2.2.3.4 Source Resetting

Source reset refers to directly resetting the parameters of the adapted model back to, or close to, the source model. It is commonly used in continual learning to alleviate catastrophic forgetting [113]. In continual test-time adaptation, source reset is applied for a similar purpose.

COTTA [175] first applies stochastic restore to reset the model. In practice, COTTA resets the model by combining the source model θ_0 and the current model weights θ_t at time step t, using

$$\theta_{t+1} = M\theta_t + (1-M)\theta_0, \tag{2.1}$$

where M is a randomly generated mask used to reset parameters partially. The straightforward strategy effectively mitigates forgetting in the short run. However, over a longer period, the strategy is ineffective [139]. ERSK [127] further adds domain shift detection by calculating the KL divergence between the running statistics in the normalisation layer up to the current time step. If the divergence is large, the model is reset. The strat-

egy avoids the unnecessary target-specific knowledge dump in COTTA. RDUMB [139] proposed a more straightforward but effective resetting method. It resets the model after a fixed number of steps. Although the accuracy drops slightly after resetting, as the target-specific knowledge is dumped, it maintains its performance over the long term by preventing forgetting. In the third work, EBaR, an elastic reset methodology is designed to preserve the source knowledge effectively without discarding the useful target knowledge.

2.2.3.5 Adapter Tuning

Adapter tuning [191] is a technique widely used with pre-trained vision models to improve downstream task performance with minimal modification. It typically introduces a lightweight module, called an adapter, which is trained specifically for the downstream task. In CoTTA, the adapter is learned during test time in a self-supervised manner. One key advantage of using adapters in CoTTA is that the source model remains unchanged, thereby preserving domain-shared knowledge from the source. As a result, catastrophic forgetting is mitigated. Also, the prompt tuning process often involves extra prompt structures [48], which are similar to adapter-based methods.

VIDA [106] introduces a Visual Domain Adapter (VDA) to separate domain-specific and domain-shared knowledge. VDA consists of a high-rank adapter that captures target-specific knowledge and a low-rank adapter that preserves shared, previously learned knowledge. A homeostatic strategy combines these two sources, enabling adaptation to new domains while preventing catastrophic forgetting. Similar to VIDA, BECOTTA [92] proposes an expert-blending approach using a Mixture-of-Domain Experts (MoDE) with a low-rank adapter. It deploys multiple lightweight adapter modules (experts) to capture target-related knowledge without affecting source knowledge. Unlike the works above, CMAE [105] introduces an autoencoder as an adapter to facilitate adaptation. CMAE masks parts of the target input using distribution-aware masking and enforces consistency between predictions on masked and original inputs. It then uses a lightweight autoencoder to reconstruct invariant features from the masked tokens, encouraging the model to learn robust, domain-relevant features. VDP [48] learns a mask-style prompt in CNNs for the target domain and a domain-agnostic prompt for shared knowledge while keeping the source fixed. During tuning, the learnable prompts are concatenated to the convolutional feature map. In the second work, SoTa-DiT, prompt-style adapters, named source and target prompts, are introduced for efficient target-specific tuning.

2.2.4 Target Adaptation Methods

2.2.4.1 Data Augmentation

Standard augmentation applies techniques such as cropping, flipping, blurring, and Gaussian noise. PAD [186] generates several augmented versions of each test image with different standard augmentation methods and uses voting on pseudo-labels to learn more general and robust target knowledge. Similarly, TTPR [45] applies multiple random augmentations to each test sample and minimises the KL-divergence to ensure consistency. This strategy is widely used in other CoTTA methods to enhance the generalisation of target-related, domain-specific knowledge. For example, in ensemble and distillation-based methods such as COTTA [175] and its successors like RMT [38], different standard augmentations are applied to a single sample to generate multiple augmented views. The predictions for these views are ensembled in the feature or label space to tune the student model.

Adaptive augmentation adjusts the data augmentation strategy to suit the target domain, addressing the issue that fixed standard augmentations may change the semantic meanings and increase learning difficulty. TESLA [165] uses adversarial augmentation to learn an augmentation policy that generates adversarially perturbed versions of target inputs, pushing them toward high-uncertainty regions and reducing learning difficulty. TIPI [124] identifies a set of input transformations that simulate likely target-domain shifts while preserving semantic meaning. At test time, it ensures that the prediction for the same sample remains semantically invariant under these transformations.

In this thesis, the standard augmentation is introduced to increase the variety of the target data for more robust test-time tuning.

2.2.4.2 Pseudo-Label Learning

Pseudo-label learning assigns labels to unlabelled data based on high model prediction confidence for certain classes. This technique is widely used in domain adaptation tasks such as semi-supervised domain adaptation [7] and unsupervised domain adaptation. Pseudo-label learning in CoTTA involves four key steps: 1) label assignment, 2) label refinement, 3) label filtering, and 4) learning with labels. In this section, we focus on the first three steps, as the third is closely related to the optimisation objective design and is discussed in detail in the next section.

Label assignment methods focus on generating and assigning more reliable pseudo-labels to each unlabelled test sample. PAD[186] uses multiple augmentations of each test

sample to vote on a pseudo-label, improving label reliability. CoTTA [175], MEMO [211], and TESLA [165] reduce pseudo-label noise by averaging the predictions across multiple augmentation views. This multi-augmentation averaging strategy is widely applied in subsequent CoTTA works. Beyond multi-augmentation aggregation, ECL [206] proposes learning from complementary labels.

Label refinement methods focus on adjusting the generated pseudo-labels to improve their quality. The most common approach for pseudo-label refinement is memory bank-based methods. These methods collect sample features across different mini-batches and apply techniques such as graph-based label transfer to propagate labels across similar samples. AdaContrast [19] refines pseudo-labels using a memory bank of target features with a nearest-neighbour voting scheme. TeSLA [165] adopts the same strategy as AdaContrast for feature refinement. TSD [178] uses a memory bank for prototype-based label refinement. It builds a memory of past target features and computes pseudo-prototypes for each class. Beyond memory bank-based methods, AR-TTA [151] incorporates source data guidance via mixup to enhance the quality of pseudo-labels. It assumes access to a small memory of labelled source examples during test time. For each incoming target sample, AR-TTA performs mixup augmentation between the unlabelled target and one or more source samples. The model is then tuned on these mixed examples using a blended label that combines the one-hot source label and the target's pseudo-label.

Label filtering methods focus on removing low-quality pseudo-labels based on indicators such as feature-space distance from past samples, high uncertainty, or abnormal back-propagation gradients. By filtering pseudo-labels, the average quality of those used for tuning is improved. NOTE [53] introduces Prediction-Balanced Reservoir Sampling (PBRS) to maintain class balance as the model processes a non-i.i.d. stream, PBRS filters and stores samples in a reservoir such that the pseudo-label class distribution remains roughly uniform. ROTTA [202] introduces a similar category-balanced buffer for non-i.i.d. data streams. EATA [128] filters out unreliable pseudo-labels using an active sample selection criterion. This criterion identifies samples with low uncertainty and rich target-related information, excluding those that are either overly complex or overly simplistic. TSD [178] filters noisy labels using an entropy filter that excludes samples with high prediction entropy and a consistency filter that drops samples whose labels are distant from the aggregated labels of their closest prototypes. SAR [129] filters out samples that would induce noisy gradient updates.

In this thesis, EBaR uses label filtering to separate low- and high-uncertainty labels, thereby reducing the influence of target noise.

2.2.4.3 Optimization Objective Designing

Optimisation objective designing aims to transfer existing knowledge to the target domain and effectively extract target-specific knowledge. There are two primary design purposes: 1) uncertainty minimisation, and 2) target knowledge generalisation.

Uncertainty minimisation focuses on reducing prediction uncertainty in the target domain. Specifically, there are two types of methods: entropy-based methods and clustering-based methods.

Entropy-based methods are the most commonly adopted for uncertainty minimisation, typically by minimising entropy loss. It increases the prediction confidence and ensures that class boundaries lie in low-density regions of the feature space [18, 184]. TENT [170] first proposed optimising the model via standard entropy minimisation for the CoTTA task. TENT adapts the model at test time by minimising entropy:

$$\mathscr{L}_{\text{Entropy}} = -\frac{1}{N}\sum_{i=1}^{N}\sum_{c} p_\theta(y=c|x_i)\log p_\theta(y=c|x_i), \tag{2.2}$$

where N is the number of samples in a mini-batch, and $p_\theta(y=c|x_i)$ is the probability that x_i belongs to class c predicted by the model. This loss encourages the predicted distribution $p_\theta(y|x_i)$ to be sharp, with one class having high confidence in its prediction. CONJ-PL [57] further proves that entropy minimisation is particularly suitable for models pre-trained with cross-entropy loss.

Clustering-based methods push samples into tighter clusters in the feature space. As samples move closer to their respective class centres while being pushed away from other centres, prediction uncertainty decreases. The typical way to achieve these is to use the class centroid or class prototype. CORE [201] calculate the averaged feature output for each class and pushes the averaged feature centres away from each other using the loss function given by:

$$\mathscr{L}_{\text{Core}} = \sum_{j=1}^{C}\sum_{j'\neq j}^{C} \mathbf{p}_j^\top \mathbf{p}_{j'}, \tag{2.3}$$

where $\mathbf{p}_j$ is the class center for class j. CFA [84] computes the class centres, the sample distribution around each centre, and the overall class-wise centres for both the limited available source data and the current test batch. These three types of centres are then aligned using a mean square distance loss.

Target knowledge generalisation focuses on extracting generalised target knowledge from unlabelled target data. Specifically, there are two main categories of approaches: contrastive-based methods and perturbation-based methods.

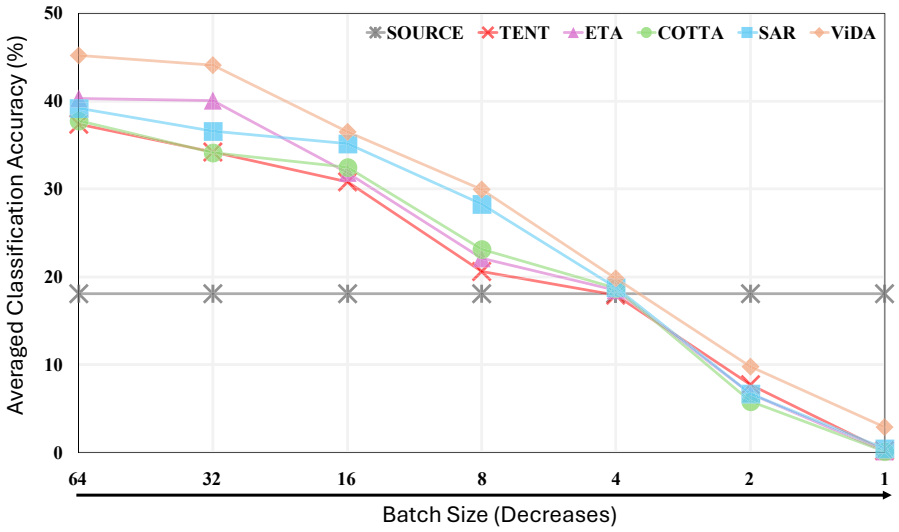


Figure 2.3: The classification accuracy (%) of the CoTTA methods under different batch size settings with ResNet-50 [155, 164] on ImageNet-C. The accuracy declines drastically following the decrease in batch size due to tuning instability.

Contrastive-based methods adopt contrastive learning strategies and contrastive loss as the objective. The contrastive loss enables the model to learn target-invariant representations from a small amount of unlabelled data within a batch [117]. AdaContrast [19] first adopts the contrastive learning loss for CoTTA. Specifically, AdaContrast employs an InfoNCE-style contrastive objective: for a sample x_i, with an augmented view x_i^+ as a positive and other samples x_j $(j \neq i)$ as negatives, the loss is given by:

$$\mathrm{L}_{\mathrm{Con}} = -\log \frac{\exp(\mathrm{sim}(z_i, z_i^+)/\tau)}{\exp(\mathrm{sim}(z_i, z_i^+)/\tau) + \sum_{j \neq i} \exp(\mathrm{sim}(z_i, z_j)/\tau)}, \tag{2.4}$$

where $z = f_\theta(x)$ are the normalized feature embeddings and $\mathrm{sim}(u, v)$ denotes cosine similarity; τ is a temperature. Following AdaContrast, [38] and [198] adopt the same contrastive learning loss.

Perturbation-based methods introduce perturbations to the model and expect its predictions to remain stable under such perturbations. SAR [129] incorporates sharpness-aware minimisation through perturbation. Inspired by SAM [46], SAR seeks model weights that perform well within a neighbourhood around θ, rather than optimising for a specific point estimate. Formally, for a given test sample (or batch) x, the optimisation objective of SAR is defined as:

$$\min_{\theta} \max_{\|\epsilon\| \leq \rho} H(p_{\theta+\epsilon}(y \mid x)), \tag{2.5}$$

where the inner maximisation identifies an adversarial weight perturbation ϵ.

In this thesis, both entropy loss and contrastive loss are adopted for SoTa-DiT and EBaR. Our refined entropy loss effectively minimized prediction uncertainty, while the contrastive loss effectively learned invariance within continually changing target data.

2.2.4.4 Gradient Adjusting

Gradient adjustment techniques aim to refine gradients during optimisation to enhance knowledge transfer to the target domain.

One approach to gradient adjustment is holistic gradient modulation, which involves introducing an additional module to adjust gradients globally. SLWI [93] employs a Bayesian filtering approach to regulate gradient updates. It treats model weights as latent variables that remain stationary unless compelling evidence suggests otherwise. Each incoming test sample updates the weight posterior through a Bayesian update, replacing raw stochastic gradient descent (SGD) steps. Another approach is to personalise the gradient for different layers or individual parameters. DAT [125] dynamically partitions parameters into domain-specific parameters and task-relevant parameters. It identifies parameters sensitive to domain shift as the domain-specific parameters by observing batch updates; if a gradient exceeds a threshold, the parameter is classified as DSP. During adaptation, only DSPs are updated. AWN [133] adjusts gradient magnitudes per layer using an information-theoretic importance measure based on the Fisher Information Matrix.

2.2.5 Research Gaps

Recent methods primarily focus on either mitigating catastrophic forgetting or enhancing target adaptation, with only a few approaches addressing both challenges simultaneously. Among them, VDP [48] is the only method that explicitly employs prompt tuning to disentangle source and target knowledge, applying distinct learning strategies to separate prompts to achieve both objectives. However, the VDP framework is limited in its applicability, as it is explicitly designed for CNN-based models and does not extend to transformer architectures.

To address this limitation, a novel transformer-based prompt tuning method is proposed in Chapter 4, aiming to provide an effective and generalisable solution for transformer baselines.

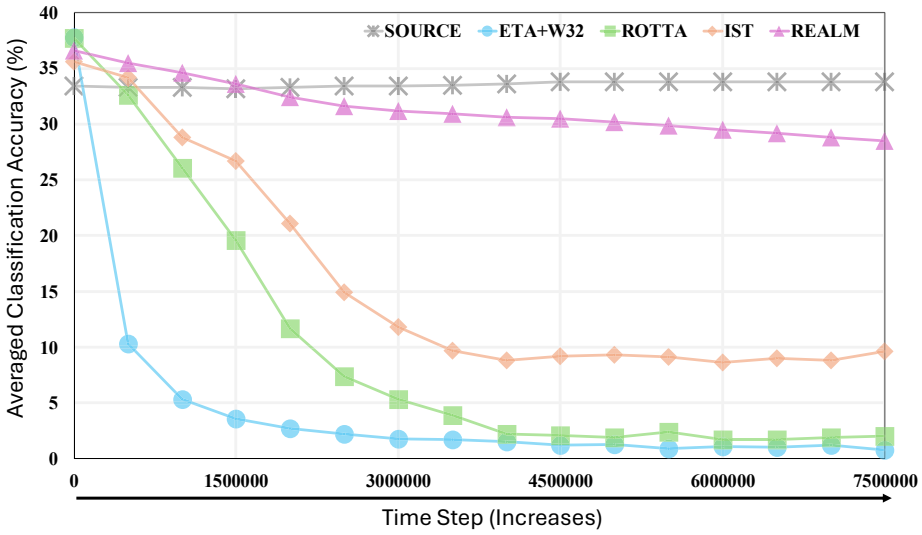


Figure 2.4: The classification accuracy (%) of the STTA methods under the S-CoTTA setting on the CCC dataset with the ResNet-50 baseline. We apply a moving window with size 32 for ETA, denoted as 'ETA+W32'. The accuracy declines following the increase in the time step due to catastrophic forgetting.

Moreover, [182] finds that when using a vision transformer backbone, recent CoTTA methods tend to collapse as the test batch size decreases. As shown in Figure 2.3, an observational experiment further reveals that this collapse is even more severe for CNN-based models [64]. In response to this issue, a novel setting, *Single-Sample Continual Test-Time Adaptation* (S-CoTTA), is introduced in Chapter 5, where the test-time batch size is set to one.

Additionally, another observation indicates that the performance of existing single-sample test-time adaptation (STTA) methods deteriorates over an extended time. To address these limitations, this thesis proposes a memory-efficient buffer and elastic resetting method in Chapter 5, designed to jointly: (1) stabilise the adaptation process under the single-sample setting, and (2) prevent catastrophic forgetting during long-term adaptation.

CHAPTER 3

CROSS-DOMAIN FEW-SHOT IMAGE CLASSIFICATION VIA PROD: PROMPTING-TO-DISENTANGLE

3.1 Introduction

Few-shot image classification aims to use limited support samples to transfer the classifier from base training classes to novel test classes [44, 150, 158, 168, 169], which meets the requirement in application scenarios when labelled data is scarce and the novel classes introduce label shift. In this thesis, we focus on the few-shot image classification in the cross-domain learning setting, where there is a domain gap between the training set and the test set. This train-to-test domain gap hinders knowledge transfer between training and test data, significantly compromising classification accuracy [23, 60]. Here, we tackle the cross-domain few-shot image classification (CDFSIC) task.

Generally, there are two approaches for mitigating the domain gap, *i.e.*, domain generalisation, and domain adaptation. Domain generalisation improves the inherent generalisation ability of the learned feature and directly applies it to novel domains without further tuning. In contrast, the domain adaptation uses samples from the novel domain to fine-tune the already-learned feature. For few-shot image classification, the domain generalisation approach [72, 99, 166, 214] is more explored than the domain adaptation approach [60], since limited support samples hardly provide reliable clues for

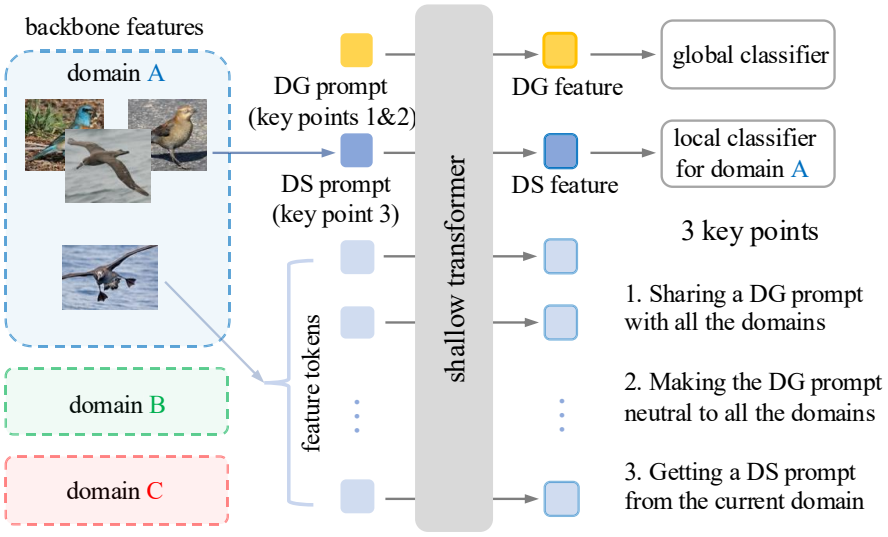


Figure 3.1: ProD flattens the CNN backbone feature into feature tokens and concatenates them with the DG and DS prompt. The DG prompt is learnable and shared by all the training domains for general knowledge. In contrast, the DS prompt is generated from the features of samples in the same domain as the feature tokens and thus can capture novel test domain knowledge from the support images during the test. The output of the DG/DS prompt is supervised with a global/local classification head, respectively, during training and concatenated as the final representation for inference..

domain adaptation.

To this end, we propose a prompting-to-disentangle (ProD) method through a novel exploration with the prompting mechanism. The prompting technique was first introduced in natural language processing and has become popular in computer vision [78, 220]. It aims to switch the transformer to different mapping functions without changing its parameters by using different prompts to condition (impact) the transformer. Compared with prior prompting techniques, our exploration is novel. With a single transformer, we simultaneously use two prompts to extract the domain-general (DG) knowledge and the domain-specific (DS) knowledge in parallel. Therefore, these two prompts switch a single transformer between two different outputs simultaneously, *i.e.*, DG and DS knowledge, yielding the so-called Prompting-to-Disentangle.

In ProD, both the DG and DS knowledge are beneficial, contrary to prior works [72, 95, 103] where the DS knowledge is considered harmful and discarded. The reason is that in ProD, the DS knowledge is not bound by the already-seen training domains. Instead, it can on-the-fly capture the novel domain knowledge from support samples

through the prompting mechanism (as explained in the third benefit below). Therefore, ProD benefits from the DS knowledge of the novel test domain. Specifically, as illustrated in Figure 3.1, ProD adopts the popular multi-domain training scheme [72, 166] and uses a Convolutional Neural Network (ResNet-10 [64]) to extract the backbone feature. Afterwards, ProD flattens the backbone feature into multiple feature tokens, concatenates the feature tokens with a DS and a DG prompt, and feeds them into a lightweight transformer. The DG prompt is learnable and shared by all the training domains, while the DS prompt is generated with backbone features from the domain-of-interest (*i.e.*, the domain of the feature tokens) on the fly. In ProD, there are three key points for mitigating the domain gap:

1) Sharing a single prompt for all the training domains benefits cross-domain generalisation. In other words, we need no special design to obtain a DG prompt but only to share a single prompt with multiple domains. During training, the output state of the DG prompt (*i.e.* the DG feature in Figure 3.1) is fed into a global classifier that contains the categories from all the training domains. Inference with the DG feature improves classification accuracy.

2) The DG prompt can be further reinforced by making it neutral towards all the training domains. To this end, we enforce a simple constraint: the learned DG prompt should have identical (or close) distance toward all the training domains. This constraint reduces the bias toward any domain and enriches the domain-general knowledge, bringing another round of improvement.

3) The DS prompt can capture the DS knowledge from the domain-of-interest on the fly and thus makes the DS knowledge beneficial. Specifically, during training, given an input, we use features from the same domain to generate a DS prompt. Correspondingly, the knowledge in the DS prompt is from the input domain specifically rather than from all the training domains. Moreover, the output DS feature is supervised by a local classifier, which contains only the categories in the current domain and thus avoids cross-domain interference. In the inference phase, we duplicate the DS prompt generation procedure onto the test domain, *i.e.*, generating the DS prompt from more support samples. Therefore, although the model remains unchanged, the on-the-fly DS prompt modifies the context of the model input and dynamically conditions the output to the test domain. Such a prompting and conditioning effect can be viewed as a test-time adaptation without fine-tuning the model.

ProD concatenates the DG and DS features as the final representation for inference, therefore integrating the benefits of good generalisation and fast adaptation. Conse-

quently, ProD effectively mitigates the domain gap and improves cross-domain few-shot classification. For example, on CUB, ProD improves the 5-way 5-shot recognition accuracy from 73.56% to 79.19% on the CUB dataset, setting a new SOTA.

The contributions of ProD can be summarized as follows:

• We propose a Prompting-to-Disentangling (ProD) method for cross-domain few-shot image classification. ProD disentangles the domain-general (DG) and domain-specific (DS) knowledge through a novel exploration of the prompting mechanism.

• For the DG knowledge, we show that sharing and neutralising a DG prompt for all the training domains benefits cross-domain generalisation. For the DS knowledge, we condition the model to the novel test domain through a DS prompt generated on-the-fly to replace fine-tuning.

• We conduct extensive experiments to validate the effectiveness of ProD. Ablation studies show that both the DG and DS prompt in ProD are effective.

3.2 Visual Prompting Mechanism

The prompting technique [13] was first introduced in natural language processing. It modifies the pre-trained language model for different downstream tasks by changing the prompt instead of tuning the deep model. Recently, the prompting mechanism has been applied in vision tasks for efficient fine-tuning [78]. Compared with the existing methods [107], our ProD has close connections and significant differences. Similar to prior works, in ProD, the prompt changes the mapping function of the deep model by modifying the context of the model input and does not change the model parameters. Still, there are two significant differences regarding training and inference. 1) *Training*: recent prompting techniques usually require a pre-trained model. Then, the prompts are injected and tuned for novel downstream tasks. When tuning the prompt, the pre-trained model is frozen. In contrast, in the proposed ProD, the "base model" (the CNN and the transformer head) and the prompts are trained simultaneously from scratch in an end-to-end manner. 2) *Inference*: Different prompts are usually employed separately in recent popular prompting techniques. In contrast, ProD simultaneously injects two prompts to activate two different knowledge in parallel.

3.3 Methodology

3.3.1 Problem Formulation

ProD adopts the popular multi-domain training paradigm [72, 82, 95, 109] for solving the CDFSIC task. Specifically, we use a group of training datasets corresponding to different domains $\mathscr{D} = \{\mathtt{D}_0, \mathtt{D}_1, ..., \mathtt{D}_N\}$. In each training iteration, images are randomly sampled from a dataset, and a few-shot learning episode is conducted on the corresponding domain.

The test dataset is from a novel domain $\mathtt{D}_t$ and contains images with disjoint labels. When testing, only a few labelled samples are provided for fine-tuning the model. Each testing episode performs a C-way K-shot task by randomly sampling a support set and a query set from the test domain $\mathtt{D}_t$. The support set consists of $C \times K$ (C is the number of classes, and each class has K samples) samples, and the query set consists of multiple unlabelled images from these C classes.

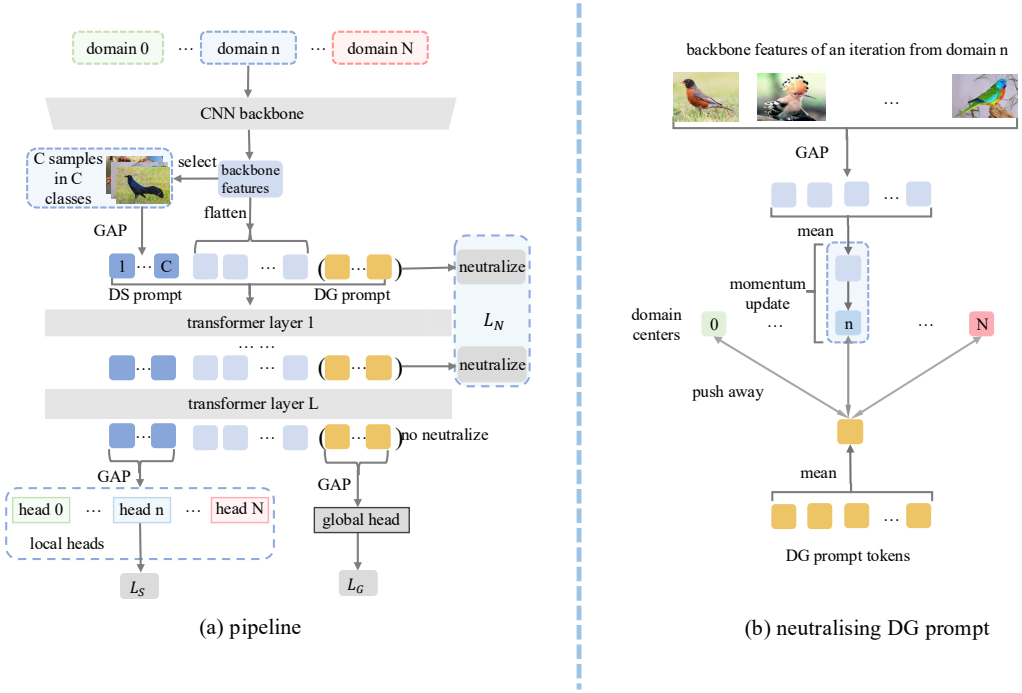
3.3.2 Overall Architecture

The overall pipeline of ProD is illustrated in Figure 3.2 (a). In each training iteration, a training domain $\mathtt{D}_n \in \mathscr{D}$ is randomly selected, and multiple images $\{x_i^n\}$ are sampled. The superscript n indicates the n-th domain. ProD feeds these images into a CNN backbone, denoted as $f()$, to produce their backbone features. Correspondingly, for each image x (the superscript and upper-script are omitted), its backbone feature is a convolutional feature map $f(x) \in \mathbb{R}^{D \times H \times W}$ (D, H, W are the dimension, height, and width).

Given a backbone feature $f(x)$, ProD flattens it into a feature sequence consisting of $H \times W$ tokens (each token is D-dimensional), *i.e.*, $\mathbf{F} \in \mathbb{R}^{(HW) \times D}$. These feature tokens are then concatenated with a DG prompt ($\mathbf{G}$) and a DS prompt ($\mathbf{S}$). Since the concatenated tokens are to be input into the multi-block transformer, we add a superscript "0" to indicate their position (*i.e.*, $\mathbf{F}^0$, $\mathbf{G}^0$ and $\mathbf{S}^0$). The DG prompt $\mathbf{G}^0$ consists of multiple tokens and is learnable (Sec.3.3.3). The DS prompt $\mathbf{S}^0$ is generated from some other backbone features in the current domain $\mathtt{D}^n$ during training (Sec.3.3.4). During testing, $\mathbf{S}^0$ is generated from the support samples in $\mathtt{D}_t$ to capture the novel domain knowledge. ProD feeds the concatenated tokens into a transformer with L blocks (we empirically set $L = 2$), which is formulated as:

$$[\mathbf{F}^l, \mathbf{G}^l, \mathbf{S}^l] = B_l([\mathbf{F}^{l-1}, \mathbf{G}^{l-1}, \mathbf{S}^{l-1}]), \tag{3.1}$$

where B_l ($l = 1, 2, \cdots, L$) is the l-th transformer block, [ ] is the concatenation operation.



(a) pipeline

(b) neutralising DG prompt

Figure 3.2: The architecture of ProD. Fig.(a) and Fig.(b) respectively depict the overall pipeline and the DG prompt neutralisation. GAP represents global average pooling. As shown in (a), ProD first extracts the backbone feature with a CNN backbone. Then, it flattens the backbone feature into feature tokens and concatenates them with a learnable DG prompt and an on-the-fly DS prompt. The outputs of the DG / DS prompt are fed into a global/local classification head, respectively. In (b), we maintain the domain centres using a momentum update. The DG prompt, in all the transformer blocks except the final output state, is pushed away from the domain centres for neutralisation.

During training, the output state of GD and GS prompts (*i.e.*, $\mathbf{G}^L$ and $\mathbf{S}^L$) are fed into a global and local classification head, respectively. The global classification head $\mathtt{H}^{\mathtt{global}}$ contains the holistic classes from all the domains $\mathscr{D} = \{\mathtt{D}_0, \mathtt{D}_1, \cdots, \mathtt{D}_N\}$. In contrast, the local classification head $\mathtt{H}^n$ only contains the classes from the current domain $\mathtt{D}_n$ ($n \in 1, 2, \cdots, N$).

During testing, the concatenation of $\mathbf{G}^L$ and $\mathbf{S}^L$ is used as the final representation, and the feature tokens $\mathbf{F}^l$ are discarded. $\mathbf{G}^L$ contains domain-general knowledge, while $\mathbf{S}^L$ contains domain-specific knowledge conditioned by the novel test domain. Combining these two features brings complementary benefits for cross-domain evaluation (Sec.3.4.3).

The following Sec.3.3.3 elaborates on the DG prompt, neutralising the DG prompt for better domain generalisation and the corresponding global classification head. Sec.3.3.4 elaborates on the DS prompt and the corresponding local classification head.

3.3.3 Domain-General Prompt for Domain-General Feature

3.3.3.1 DG Prompt and Global Classification

Domain-General prompt $\mathbf{G}^0$ contains multiple trainable tokens. When the feature tokens and the DG tokens ($[\mathbf{F}^0, \mathbf{G}^0]$) proceed in the transformer, they interact with each other through the attention mechanism. After L transformer blocks (Eqn.3.1), we consider the output $\mathbf{G}^L$ as containing domain-general knowledge and use a global classification head to supervise $\mathbf{G}^L$, which is formulated as:

$$\mathscr{L}_G = -\log \frac{\exp(\mathbf{w}_y \cdot \overline{\mathbf{G}^L})}{\sum_j \exp(\mathbf{w}_j \cdot \overline{\mathbf{G}^L})}, \tag{3.2}$$

where $\mathbf{w}_j$ enumerates all the weight vectors in the global classification head. $\mathbf{w}_y$ is the weight vector of the ground-truth class. " $\cdot$ " is the inner product operation. " $\overline{\ }$ " is the average pooling operation, through which multiple tokens in $\mathbf{G}^L$ are pooled into a vector $\overline{\mathbf{G}^L}$. This global classification head covers all the classes from the entire training domains $\mathscr{D} = \{\mathtt{D}_1, \mathtt{D}_2, \cdots, \mathtt{D}_N\}$.

Empirically, we find that adding DG prompt brings considerable improvement (*e.g.*, in Sec.3.4.3, +1.82 top-1 accuracy on CUB) over the "CNN+Transformer" baseline. Analytically, it is because the DG prompt brings additional input tokens shared by all the training domains. Intuitively, if the number of DG prompt tokens is huge, different input images have almost the same representation. Under this extreme assumption, the deep representation has minimal domain bias. However, it lacks basic discriminative ability and might not function properly. Therefore, the number of DS prompt tokens matters and is empirically set to 5, as illustrated in Sec.3.4.3.

3.3.3.2 Neutralizing DG Prompt

We further neutralise the DG prompt to reinforce its generalisation ability. The intuition is that if the DG prompt has no bias to any training domains $\mathscr{D}$, it can capture more general knowledge and further benefit cross-domain generalisation. To this end, given a DG prompt, we measure its domain bias through the cosine similarity towards different training domain centres, as illustrated in Figure 3.2 (b). A domain centre is the averaged feature of all the samples in the corresponding domain, which can be approximated online by a momentum update:

$$\mathbf{p}^n \leftarrow \lambda \mathbf{p}^n + (1-\lambda)\frac{\sum_{i=1}^{B}\overline{f(x_i^n)}}{B}, \tag{3.3}$$

where the superscript n indicates the n-th training domain, λ is the momentum rate, B is the batch size, $\overline{f(x_i^n)}$ is a vector pooled from the backbone feature through the average pooling " $\overline{\ }$ ".

Given the DG prompt, the neutralising loss minimises its cosine similarity to all the domain centres, which is formulated as:

$$\mathscr{L}_N = \frac{1}{N}\sum_{n=1}^{N}(|\frac{\overline{\mathbf{G}}\cdot\mathbf{p}^n}{\|\overline{\mathbf{G}}\|\|\mathbf{p}^n\|}|). \tag{3.4}$$

The above neutralizing loss (Eqn.3.4) is performed to G^l ($l = 0, 1, \cdots, L-1$). We choose not to neutralise G^L because it conflicts with the global classification loss (which is also on G^L).

3.3.4 Domain-Specific Prompt for Domain-Specific Feature

3.3.4.1 Domain-Specific Prompt

The DS prompt $\mathbf{S}^0$ is on-the-fly generated from some random backbone features $f(x)$ in the current domain through average pooling, as illustrated in Figure 3.2 (a). Specifically, during training, ProD chooses C training classes, randomly samples 1 backbone feature $f(x)$ from each class, and uses the average-pooled vector $\overline{f(x)}$ as a corresponding token. Consequently, $\mathbf{S}^0$ contains C tokens, *i.e.*, $\mathbf{S}^0 \in \mathbb{R}^{C\times D}$. During testing ($C$-way K-shot), ProD duplicates the generation procedure onto the novel testing domain, *i.e.*, using C support samples (1 from each class) to derive the DS prompt.

A side-effect of the DS prompt is that each DS token contains the underlying domain knowledge and the class information from the initialisation. While the DS prompt intends to inject the domain knowledge, the class information is NOT desired. It might become a distraction (because among all the C DS tokens, $C-1$ tokens belong to different classes as the query image).

3.3.4.2 Local Classification Head

To suppress the undesired class information, we design a "changing identity" objective: after the DS prompt proceeds in the transformer along with the feature tokens $\mathbf{F}$, each

DS token should lose its own class identity and change the identity to the same as $\mathbf{F}$. To this end, ProD feeds the output state of the DS prompt (after average pooling), *i.e.,* $\overline{\mathbf{S}^L}$ into a local classification head. The local classifier only covers the training classes in the current n-th domain, which is formulated as:

$$\mathscr{L}_S = -\log \frac{\exp(\mathbf{u}_y^n \cdot \overline{\mathbf{S}^L})}{\sum_j \exp(\mathbf{u}_j^n \cdot \overline{\mathbf{S}^L})}, \tag{3.5}$$

where $\mathbf{u}_j^n$ enumerates all the weight vectors in the n-th local classification head, u_y^n is the weight vector for the ground-truth category of the input feature tokens $\mathbf{F}^0$, rather than any ground-truth categories for generating the GS prompt.

The reason for using the local classification head instead of a global head is: in a global classification head, $\overline{\mathbf{S}^L}$ (from a specific domain $\mathtt{D}_n$) interacts with all the weight vectors from the entire training domains $\mathscr{D} = \{\mathtt{D}_1, \mathtt{D}_2, \cdots, \mathtt{D}_N\}$. The global interaction will propagate domain knowledge from other domains ($\mathtt{D}_{j \neq n}$) to $\overline{\mathbf{S}^L}$ and thus blur its domain-specific knowledge.

The overall loss for ProD is calculated as:

$$\mathscr{L} = \mathscr{L}_G + \alpha \mathscr{L}_N + \beta \mathscr{L}_S, \tag{3.6}$$

where $\mathscr{L}_G$, $\mathscr{L}_N$, $\mathscr{L}_S$ are the global classification loss (Eqn.3.2), neutralizing loss (Eqn.3.4), and the local classification loss (Eqn.3.5), respectively. α and β are the balancing hyper-parameters.

3.4 Experiments

3.4.1 Settings

Datasets. Following the popular multi-domain training scheme, we use miniImageNet [148] and four fine-grained datasets, *i.e.*, CUB [11], Cars [86], Plantae [167] and Places [217]. We adopt the leave-one-out setting, *i.e.*, choosing one fine-grained dataset for inference and using the other three fine-grained datasets along with miniImageNet for training.

3.4.1.1 Baseline

The baseline of ProD consists of CNN (ResNet-10 [64]) and a lightweight Transformer head. Without DG or DS prompt, this "CNN+Transformer" baseline achieves 72.32% 5-way 5-shot accuracy in CUB and outperforms the popular pure CNN baseline (68.98%) by

+3.34%, as detailed in Sec.3.4.5. We note that: 1) a strong "CNN+Transformer" baseline only contributes a small portion to the superiority of ProD because ProD further improves the baseline by a large margin (e.g., +6.87% on CUB), and 2) adding the transformer head increases the model size (5.3M → 8.5M, as discussed in Sec.3.4.5) but is still efficient (smaller than ResNet-18 but achieves higher accuracy).

3.4.1.2 Inference

In the C-way K-shot testing phase, the transformer and the CNN backbone are both frozen. Then, we randomly select C support samples (each from a class) and use their backbone features to generate the DS prompt on the fly. Finally, the DS and DG prompt outputs are concatenated as the feature representation. Finally, we use the support features to learn a new linear classification head and then use the new head to classify the query samples, consistent with the standard few-shot classification pipeline.

3.4.1.3 Network Architecture

ResNet10 [64] is the backbone of our ProD. The input image size is $3 \times 224 \times 224$. The output of ResNet10 before the classification head serves as a backbone feature, whose size is $7 \times 7 \times 512$ and is reshaped into 49×512 (49 feature tokens) later to be concatenated with the prompts. The architecture of the transformer layers following the backbone is the same as the transformer unit in [40] (standard ViT transformer unit) with one transformer layer and one MLP layer. The sizes of the DS prompt and DG prompts are set as 5, and their embedding dimensions are set as 512, the same as the backbone feature. Concatenated with the prompts, the shape of the input for the transformer layers is 59×512. The hidden embedding size for the transformer is 512, and the number of transformer heads is 8. After the transformer layer, the size of the output vector is 59×512. Finally, the outputs of DS and DG prompts (both with size 5×512) are averaged respectively to classify the image.

3.4.1.4 Multi-domain Training Scheme

We use miniImageNet [148] as our base source dataset since it contains images of immense variety. Four fine-grained cross-domain datasets, including CUB [11], Cars [86], Plantae [167] and Places [217], are selected as other source datasets or target dataset. When one of the four fine-grained cross-domain datasets serves as a target domain dataset for inference, the other three serve as source domain datasets for training.

In each training batch, half of the samples are from the base dataset, miniImageNet and the other half from a cross-domain dataset. Thus, different mini-batches contain images from different domains, and the three domains from the four datasets (CUB, Cars, Plantae, Places) serve as the source domain of the mini-batch in turn. To be compatible with the DS prompt, "2k" samples are selected in each training batch, where "k" is the number of categories in a mini-batch. For each category, two samples are selected. For instance, when the batch size is 64, we have $k = 32$. When inference, only one domain other than the three domains selected as the source is sampled for support and query sets.

3.4.1.5 Evaluation Protocol

Following [23, 60], we evaluate our model by sampling 600 independent 5-way few-shot classifications on the four cross-domain datasets. In each sampled test, K images from 5 novel categories are selected as a support set whose labels are available for training or fine-tuning. 15 images from the 5 novel categories are selected as a query set whose labels are not available and cannot be used to train or fine-tune the model. Following the standard setting [23, 60], we let $K = 1, 5$. The last linear classification head is re-trained from scratch based on the support set, and the rest of the parameters are frozen for inference. Then the model performance is evaluated on the query set with all the parameters frozen. For each independent test, the linear classification head is re-trained. Statistical information of query images is only used for batch normalization [23, 60]. The model is evaluated 600 times in each experiment, and the average accuracy with a 95% confidence interval, beginning with $\pm$, is reported as the model performance.

3.4.1.6 Default Hyper-Parameter Setting

The default hyper-parameter setting is shown in Table 3.1. The model is trained with a batch size of 64 for 500 epochs. The loss weight parameter α is set as 1, β is set as 1, and the domain centre momentum update rate λ is set as 0.9. Transformer depth is set as 2, and both DS and DG prompt sizes are 5. The model is optimised with adaptive moment estimation (ADAM), with a learning rate of 10^{-2} and momentum of 0.9. When inference, the linear classifier is optimised with stochastic gradient descent (SGD), with a learning rate of 10^{-2} and trained for 100 epochs.

Table 3.1: Default hyper-parameter setting.

parameter	value
α	1
β	1
λ	0.9
transformer head	8
transformer hidden embedding	512
transformer depth	2
learning rate train	10^{-2}
learning rate inference	10^{-2}
training batch size	64
training epoch	500
inference re-train epoch	100

Table 3.2: Comparison between ProD and SOTA methods on 5-way 1-shot task. Prod surpasses recent methods on four datasets: CUB, CARS, Plantae, and Places by clear margins.

Methods	Datasets			
	CUB	CARS	Plantae	Places
RelationNet [158]	35.21 ± 0.46	30.12 ± 0.49	31.99 ± 0.51	49.79 ± 0.57
MatchingNet [168]	42.28 ± 0.61	28.91 ± 0.56	33.02 ± 0.56	48.53 ± 0.62
RelationNet+LFT [166]	48.10 ± 0.62	32.26 ± 0.58	35.21 ± 0.59	51.02 ± 0.56
MatchingNet+LFT [166]	43.38 ± 0.58	30.68 ± 0.59	35.10 ± 0.54	52.63 ± 0.55
RelationNet+ATA [172]	48.49 ± 0.61	31.92 ± 0.58	33.62 ± 0.49	51.00 ± 0.50
DSL [72]	50.15 ± 0.80	37.13 ± 0.69	41.17 ± 0.80	53.16 ± 0.88
Baseline	48.56 ± 0.72	33.15 ± 0.64	37.94 ± 0.71	49.81 ± 0.69
ProD	$\mathbf{53.97 \pm 0.71}$	$\mathbf{38.02 \pm 0.63}$	$\mathbf{42.86 \pm 0.59}$	$\mathbf{53.92 \pm 0.72}$

3.4.2 Effectiveness of ProD

We compare the proposed ProD with the baseline and SOTA methods in Table 3.2 (5-way 1-shot) and Table 3.3 (5-way 5-shot). For a fair comparison, all the competing methods use the multi-domain training scheme, which is usually better than the single-domain counterpart. We draw two observations:

First, ProD improves the "CNN+Transformer" baseline by a large margin. For example, under the 5-way 5-shot setting, ProD increases the accuracy by +6.87%, +6.32%, +5.77%, +5.87% on CUB, CARS, Plantae, Places, respectively. ProD only adds ten prompt tokens (5 DG tokens + 5 DS tokens) over the baseline and incurs minimal computa-

Table 3.3: Comparison between ProD and SOTA methods on 5-way 5-shot task. Prod surpasses recent methods on four datasets: CUB, CARS, Plantae, and Places by clear margins.

Methods	Datasets			
	CUB	CARS	Plantae	Places
RelationNet	51.10 ± 0.62	38.26 ± 0.58	62.99 ± 0.62	46.01 ± 0.57
MatchingNet	57.21 ± 0.63	36.98 ± 0.56	62.83 ± 0.62	43.68 ± 0.55
RelationNet+LFT	65.02 ± 0.55	43.51 ± 0.51	50.48 ± 0.46	67.34 ± 0.52
MatchingNet+LFT	61.44 ± 0.56	43.12 ± 0.52	48.49 ± 0.51	65.09 ± 0.48
RelationNet+ATA	59.42 ± 0.48	42.99 ± 0.42	45.51 ± 0.51	67.10 ± 0.41
NSAE [99]	68.17 ± 0.54	54.77 ± 0.56	59.51 ± 0.55	70.93 ± 0.54
DSL	73.57 ± 0.65	58.53 ± 0.73	62.10 ± 0.75	74.10 ± 0.72
Baseline	72.32 ± 0.77	53.17 ± 0.71	60.05 ± 0.69	69.13 ± 0.60
ProD	$\mathbf{79.19 \pm 0.59}$	$\mathbf{59.49 \pm 0.68}$	$\mathbf{65.82 \pm 0.65}$	$\mathbf{75.00 \pm 0.72}$

Table 3.4: Comparison between ProD and SOTA methods under 5-way 5-shot setting on newly proposed datasets. Prod surpasses recent methods on three out of four datasets: ChestX, ISIC, and EuroSAT by clear margins.

Methods	Datasets			
	ChestX [179]	ISIC [29]	EuroSAT [67]	CropDisease [121]
Transductive Ft [61]	26.79 ± 0.42	49.68 ± 0.63	81.76 ± 0.82	90.64 ± 0.51
ConFeSS [32]	27.09 ± 0.71	48.85 ± 0.66	84.65 ± 0.58	88.88 ± 0.59
RDC-FT [97]$^{-}$	25.48 ± 0.49	49.06 ± 0.56	84.67 ± 0.59	$\mathbf{93.55 \pm 0.47}$
ProD	$\mathbf{28.79 \pm 0.41}$	$\mathbf{50.57 \pm 0.51}$	$\mathbf{85.09 \pm 0.58}$	90.41 ± 0.71

tional overhead (+2.51% extra computational overhead). The improvement validates the effectiveness of the proposed ProD.

Second, ProD achieves an accuracy on par with that of the SOTA methods. Under the 1-shot setting, ProD surpasses the strongest competitor DSL by +3.82%, +0.89%, +1.69%, +0.86% on CUB, CARS, Plantae, Places, respectively. Under the 5-shot setting, the superiority of ProD is even larger, *i.e.*, +6.87%, +0.96%, +3.72%, +0.90% higher accuracy on CUB, CARS, Plantae, Places, respectively. The superior performance proves that ProD deals with the CDFSIC task more efficiently and the knowledge disentangle scheme as a whole is effective.

Further, as shown in Table 3.4, ProD surpasses several newly proposed methods by a clear margin on ChestX, ISIC, and EuroSAT datasets. For these datasets, the experiments are conducted under a similar setting. We use miniImageNet and the four

Table 3.5: Evaluation of key components: DG prompt (DG), neutralising loss ($\mathscr{L}_N$), and DS prompt (DS). The results show that both DG and DS are effective and bring accuracy gain.

Methods	CUB	
	1-shot	5-shot
Baseline	48.56 ± 0.59	72.32 ± 0.67
Baseline + DG	51.89 ± 0.63	75.12 ± 0.69
Baseline + DS	51.48 ± 0.71	74.91 ± 0.68
Baseline + DG + DS	52.69 ± 0.66	77.63 ± 0.74
Baseline + DG + $\mathscr{L}_N$	53.08 ± 0.74	78.65 ± 0.68
Baseline + DG + DS + $\mathscr{L}_N$ (ProD)	$\mathbf{53.97 \pm 0.71}$	$\mathbf{79.19 \pm 0.59}$

datasets in Table 3.2 as the source for ProD and the counterpart methods. The results further demonstrate the superiority of ProD on the CDFSIC task, proving that ProD is generalisable to downstream domains.

3.4.3 Ablation Studies of Key Components

3.4.3.1 DG and DS prompts

Table 3.5 investigates two key components, *i.e.*, the domain-general (DG) and the domain-specific (DS) prompt through ablation on CUB. Based on the result, we draw three observations:

First, adding the DG / DS prompt independently improves the baseline (*e.g.*, +2.80% / +2.59% 5-way 5-shot accuracy). It indicates that DG and DS knowledge are both beneficial. We note that making the DS knowledge beneficial is particularly difficult in few-shot learning because the few samples are insufficient for fine-tuning the DS knowledge. Therefore, most of the prior works usually discard the DS knowledge. In contrast, ProD use the DS prompt, which facilitates the limited samples to serve as the condition on-the-fly. In this way, the DS knowledge of the novel test domain is injected naturally into the model without fine-tuning. The result indicates that this method successfully makes the DS knowledge beneficial.

Second, comparing "Basel. + DG + DS" against "Basel. + DG (or DS)", we find that combining the DG and DS prompts brings further improvement. The result proves that the DG and DS prompts are complementary, indicating that different knowledge is stored within the two prompts. Inexplicably, the result supports our knowledge-dependent hypothesis. Further, it indicates that the DG and conditioned DS knowledge achieve

Table 3.6: Comparison between the local and global classification heads on the DS prompt. The result shows that the local classification head suits the DS prompt as the global one represses the domain-specific knowledge learning.

Methods	CUB	
	1-shot	5-shot
Baseline	48.56 ± 0.59	72.32 ± 0.67
Baseline + DS (global)	50.39 ± 0.71	73.87 ± 0.66
Baseline + DS (local)	51.48 ± 0.71	74.91 ± 0.68
ProD (global)	52.08 ± 0.74	77.65 ± 0.68
ProD (local)	$\mathbf{53.97 \pm 0.71}$	$\mathbf{79.19 \pm 0.59}$

complementary benefits for the cross-domain challenge.

Third, neutralising the DG prompt is beneficial and brings another round improvement of +0.39% and +1.02% accuracy under the 1-shot and 5-shot setting, respectively. The result proves the effectiveness of our DG prompt neutralisation operation, demonstrating that pushing the DG prompt away from the source domain centres helps the prompt to extract the knowledge that is more general.

In conclusion, this experiment proves that all three key points in ProD are effective and beneficial to the CDFSIC task. Our hypothesis that knowledge disentanglement is beneficial to CDFSIC is implicitly proved.

3.4.3.2 Local Classification for DS Prompt

To train the DS prompt and learn the corresponding DS knowledge, ProD uses a local classification head that contains the class prototypes only in the current domain. In Table 3.6, a comparison experiment is conducted to validate this choice by replacing the local classification head with a global one, which is the same as the DG prompt. The result shows that the local heads cooperate better with the DS prompt. For example, global head reduces the 5-way 5-shot accuracy by +1.33% (only DS prompt) / +1.54% (full ProD with DG + DS prompt). The reason is that, inside the global classification head, the DS feature conducted by the DS prompt interacts with the prototypes across all the training domains. This interaction makes the DS prompt attention to the global knowledge across different domains, which is more general. However, this contradicts the purpose of the DS prompt, blurring the useful DS knowledge inside the prompt. As a result, using a global classification head, DS prompt cannot extract DS knowledge that is complementary to the DG knowledge, leading to a decrease in classification accuracy.

Table 3.7: Comparison between different features for inference with a complete ProD model. The result shows that both DG and DS features are beneficial for classification. In contrast, the feature token is not suitable for classification as it contains more local knowledge instead of global.

Inference Input	CUB	
	1-shot	5-shot
Feature Token	51.51 ± 0.72	76.13 ± 0.68
DG	53.01 ± 0.74	78.17 ± 0.61
DS	52.07 ± 0.69	77.64 ± 0.63
DG+DS	$\mathbf{53.97 \pm 0.71}$	$\mathbf{79.19 \pm 0.59}$
DG+DS+Feature Token	52.18 ± 0.75	78.04 ± 0.72

3.4.3.3 Choice of Inference Features

ProD provides three outputs, *i.e.*, feature tokens $\mathbf{F}^L$, DG tokens $\mathbf{G}^L$ and DS tokens $\mathbf{S}^L$. Each output is averaged into a vector through pooling layers. Table 3.7 investigates how to derive the most discriminative representation with these vectors. This ablation is based on a complete ProD, different from Table 3.5 where several key components are removed during training. We draw three observations below:

First, when each type is used alone, the "DG" and "DS" feature tokens are better than the "Feature Token". The reason is that the DG and DS prompt extract domain-related global knowledge for an image is distinctive. On the other hand, the feature token contains unnecessary local knowledge that is not beneficial to the classification task. The experiment results also validate that the DG and DS prompts effectively activate the DG and DS knowledge from the original backbone features and are thus superior.

Second, the feature token in ProD is better than the feature token in the baseline ("Basel." in Table 3.6), when they are applied for the classification task. The reason is that ProD implicitly propagates the DG and DS knowledge from the prompts to the feature token through the attention in the transformer. In this way, the feature token in ProD contains part of DS and DG knowledge, compared to the baseline model. As a result, they are more distinctive and perform better for classification.

Third, comparing two combination strategies against each other, we find that "DG+DS" is better. It demonstrates that the DG and DS prompts achieve a complementary benefit, while further adding the feature token compromises ProD. The result proves that the prompts fully extract the domain-related knowledge and no complementary knowledge is left within the feature tokens. Therefore, we use "DG+DS" as the final representation for the image and discard the feature tokens.

Methods	Datasets			
	CUB	CARS	Plantae	Places
RelationNet	51.10 ± 0.62	38.26 ± 0.58	62.99 ± 0.62	46.01 ± 0.57
RelationNet$^-$	51.02 ± 0.64	37.98 ± 0.59	62.78 ± 0.61	46.02 ± 0.56
MatchingNet	57.21 ± 0.63	36.98 ± 0.56	62.83 ± 0.62	43.68 ± 0.55
MatchingNet$^-$	56.92 ± 0.61	36.94 ± 0.54	62.51 ± 0.64	43.51 ± 0.57
RelationNet+LFT	65.02 ± 0.55	43.51 ± 0.51	50.48 ± 0.46	67.34 ± 0.52
MatchingNet+LFT	61.44 ± 0.56	43.12 ± 0.52	48.49 ± 0.51	65.09 ± 0.48
RelationNet+ATA	59.42 ± 0.48	42.99 ± 0.42	45.51 ± 0.51	67.10 ± 0.41
NSAE [99]	68.17 ± 0.54	54.77 ± 0.56	59.51 ± 0.55	70.93 ± 0.54
DSL$^-$	63.76 ± 0.60	51.21 ± 0.40	53.27 ± 0.49	66.12 ± 0.78
DSL	73.57 ± 0.65	58.53 ± 0.73	62.10 ± 0.75	74.10 ± 0.72
Baseline$^-$	70.98 ± 0.76	50.63 ± 0.72	58.25 ± 0.69	67.01 ± 0.57
Baseline	72.32 ± 0.77	53.17 ± 0.71	60.05 ± 0.69	69.13 ± 0.60
ProD$^-$	78.01 ± 0.79	57.22 ± 0.63	63.62 ± 0.68	72.43 ± 0.63
ProD	$\mathbf{79.19 \pm 0.59}$	$\mathbf{59.49 \pm 0.68}$	$\mathbf{65.82 \pm 0.65}$	$\mathbf{75.00 \pm 0.72}$

Table 3.8: Comparison with SOTA methods on 5-way 5-shot task with/ without a multi-domain training scheme. "$^-$" means the multi-domain training scheme is removed from the corresponding method. The result shows that the multi-domain training scheme is beneficial for CDFSIC.

3.4.3.4 Effect of the Multi-domain Training Scheme

We evaluate the effect of the multi-domain training scheme described in Sec.3.4.1.4. The result is shown in Table 3.8, where methods with "$^-$" are not trained with the multi-domain training scheme. From the result, we see that the multi-domain training scheme increases the classification accuracy of most methods. Specifically, for DSL, the performance significantly drops after removing the scheme since the model architecture is designed based on the multi-domain training scheme.

In ProD, removing the multi-domain training scheme is harmful to the performance. For instance, on the CUB, the 5-way 5-shot accuracy drops by -1.18%. The reason for this is twofold. First, removing the multi-domain training scheme hinders the DG prompt from learning DG knowledge. When the model is trained with single-source, the global classifier is ineffective and the DG prompt cannot capture any domain-invariance information, eventually reducing the generalisation ability. Second, removing the multi-domain training scheme prevents the DS prompt from learning adequate DS knowledge. The only-the-fly DS knowledge condition is not well trained with a single source and leads to poor DS knowledge learning. Combining the two reasons above, the image

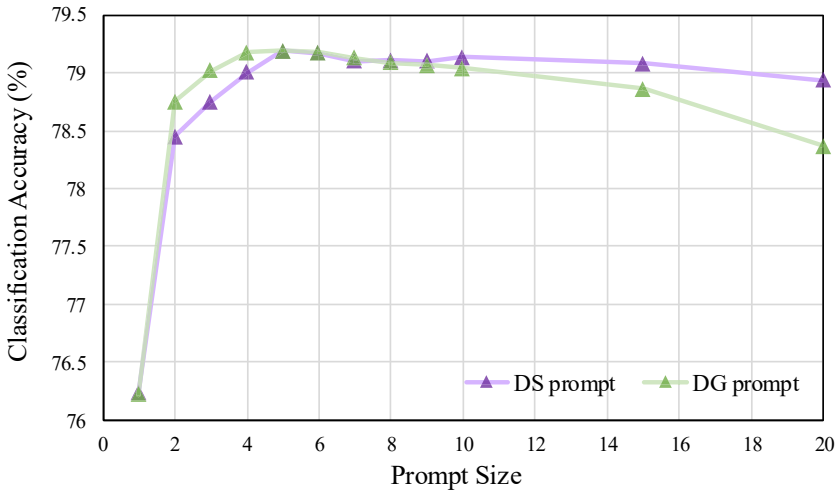


Figure 3.3: Evaluation of different sizes for DG and DS prompts with a complete ProD model. The result shows that the optimum settings for both DS and DG prompt sizes are 5.

classification accuracy drops when the multi-domain training scheme is removed.

3.4.4 Ablation Studies of Hyper-parameters

3.4.4.1 DG and DS Prompt Size

The impact of prompt size, *i.e.*, the number of tokens in the DG and DS prompts, is evaluated in this section. The results are shown in Figure 3.3, from which we draw two observations below:

First, as the DG prompt size increases, the achieved accuracy undergoes a sharp increase and a subsequent slow decrease. We infer that the reason is twofold. On the one hand, increasing the DG prompt size enhances its capability for capturing domain-general knowledge and is thus beneficial. As a result, increasing the DG size from 1 at the beginning brings improvement in classification accuracy. However, an oversized DG prompt might suppress the difference between samples because all the samples share the same DG prompt. As a result, the DG knowledge is not well concentrated and is sparsely diffused with the prompts, decreasing the discriminative ability. Therefore, we set the DG prompt size to 5, which achieves 79.19% 5-way 5-shot accuracy on CUB, for all the datasets.

Second, increasing the DS prompt size brings a similar trend of "increasing → slightly

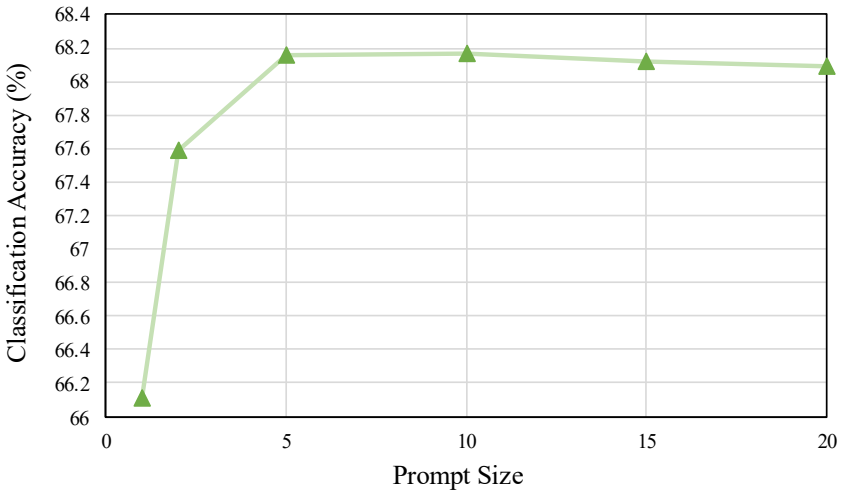


Figure 3.4: Evaluation of DS prompt sizes for 10-way 5-shot test on CUB. The result shows that under a 10-way 5-shot test, the optimum setting for the DS prompt is also 5, the same as a 5-way 5-shot test. We infer the reason is that five samples are enough to represent a test domain for CDFSIC.

decreasing" the accuracy. Moreover, different DS prompt sizes for 10-way 5-shot test, as shown in Figure 3.4. Under 10-way 5-shot test, the effect of the DS prompt size shares the same trend. The reason for this is twofold. Initially, the small DS prompt cannot include enough sample features to present a novel domain, as the size of the DS prompt is equal to the number of support sample features that can be added as a condition. As a result, the image classification accuracy increases as the DS prompt size increases when the size is small. On the other hand, the increase of DS prompt size after five does not bring benefit. The reason is that 5 support samples are enough to represent a domain under a 5 or 10-way test. Increasing the size of the prompt after that makes the DS knowledge less concentrated and harms the performance. The result also indicates that the prompt size does not have to be significantly increased with the category number C. Further, the 10-way test result is lower than the 5-way test since the 10-way task is harder. For each category, we have the same number of support samples to train the network, but the category number C is increased to 10.

3.4.4.2 Transformer Depth

The impact of transformer depth for ProD is evaluated on CUB under the 5-way 5-shot setting. All the models are the complete ProD model except for the "0 layer backbone

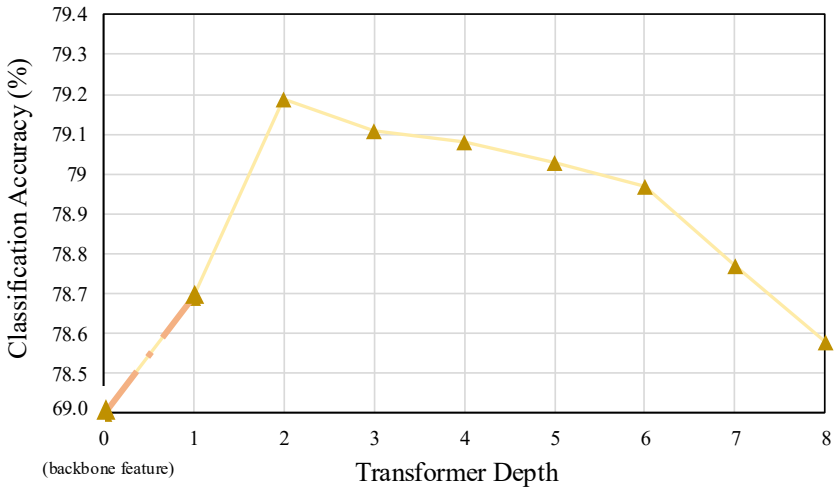


Figure 3.5: Analysis of the transformer depth (CUB dataset, 5-way 5-shot test). The ordinate is NOT linearly scaled for the "69.0→78.5" interval. The result shows that two transformer layers yield the best results.

feature". The results are shown in Figure 3.5. We draw two observations below:

First, when the transformer has only one block, it already significantly improves +9.70% over the CNN backbone feature (68.98%→78.68%). The result proves that ProD works well even with only one layer of a light transformer, as the additional computation cost within a 1-block light transformer is less than 1.6M, which is a minimal increase compared with the CNN backbone. Further, it is observed that adding a transformer without our prompting-to-disentangling mechanism only brings slight improvement. The result proves that the improvement is mainly contributed by our prompting-to-disentangling mechanism rather than the transformer itself. Combining the two points, we can conclude that ProD significantly improved the CDFSIC performance with minimal extra computational cost.

Second, when the transformer depth increases, the accuracy increases to its peak of 79.19% and gradually decreases. We infer that the reason is two-fold. On the one hand, two transformer blocks are already sufficient for depicting the required prompting-to-disentangling effect. The reason is that all the datasets used for training and testing are on a small to medium scale. Thus, no large transformer is needed to preserve all the knowledge. On the other hand, training a large transformer generally requires a large-scale dataset [40]. Under CDFSIC, the commonly applied small-scale training data is insufficient and causes over-fitting issues, especially at test time. Therefore, we set the

Table 3.9: Evaluation of different weights for neutralising loss. The result shows that 1 is the optimum setting.

Weight (α)	CUB	
	1-shot	5-shot
0.01	53.12 ± 0.72	78.75 ± 0.69
0.1	53.79 ± 0.67	79.03 ± 0.61
1	$\mathbf{53.97 \pm 0.71}$	$\mathbf{79.19 \pm 0.59}$
10	52.87 ± 0.73	78.54 ± 0.70

transformer depth to 2 as the optimised result.

3.4.4.3 Weight of Neutralizing Loss

The effect of different weights for neutralising loss: 0.01, 0.1, 1 and 10 is evaluated in this section. The result is shown in Table 3.9. When the weight is minimal, the neutralising loss is ineffective and cannot remove the domain bias within the DG prompt. As a result, the generalisation ability of the DG prompt decreases, leading to a decrease in the classification accuracy. In contrast, when the weight is too high, the effect of the neutralising loss is too strong, and the discriminative ability of the DG prompt declines as the domain-related information is overly removed. As a result, the DG prompt cannot provide a discriminative feature for classification, leading to a decline in accuracy. Based on the two observations, we set the weight α as 1 to neutralise the loss for all the other experiments in default.

3.4.5 Computational Efficiency and Cost

Table 3.10 analyzes the computational efficiency by comparing the size of the model and the accuracy achieved. Comparing "Basel. (Res10 + Trans)" against "Res10", we observe that the transformer head increases 3.2M parameters and brings +3.34% increase in classification accuracy (68.98% → 72.32%). This improvement is due to the inherent capability of the transformer. Based on the baseline, ProD further brings +6.87% increase in classification accuracy while adding only about 0.1M parameters. This indicates that the prompting-to-disentangling mechanism is the major reason for the superiority of ProD and is very efficient. Moreover, compared to the larger pure CNN model (ResNet-18), ProD (based on ResNet-10) is still more accurate while being smaller.

Regarding computational cost, the average inference time for a 5-way, 5-shot test is reported. We observe that, compared to ResNet-18, ProD achieves higher classifica-

Table 3.10: Analysis of the computational efficiency and computational cost. "Res10", "Res18" and "Trans" denote ResNet-10, ResNet-18 and the transformer head, respectively. We list the model size and the time cost per 5-way 5-shot test. The result shows that ProD outperforms other counterparts with minimal parameter increase.

Method	Size	Cost	CUB 5-shot
Res10	5.3M	42s	68.98 ± 0.81
Res18	11.7M	97s	72.39 ± 0.84
Baseline (Res10 + Trans)	8.5M	79s	72.32 ± 0.77
ProD (Res10 + Trans + Prompt)	8.6M	83s	79.19 ± 0.59

tion accuracy while incurring lower inference time. Compared to the baseline method, ProD yields an accuracy improvement of approximately +7%, at the expense of only an additional 4 seconds per test, corresponding to an increase of approximately 5% in inference time. These results indicate that ProD is an efficient CDFSIC solution, delivering substantial performance gains with minimal additional computational overhead.

3.4.6 Limitation Discussion

A limitation of the proposed approach is that, in cross-domain few-shot learning settings, ProD remains less effective when the target labels are highly fine-grained. For example, on the CARS dataset, ProD achieves a comparatively lower accuracy than competing methods, as the visual distinctions among fine-grained car categories are subtle and difficult to discriminate. A similar performance degradation is observed on the Plantae dataset, where inter-class visual differences between several categories are also minimal. These results indicate that ProD has limited capacity to capture subtle, fine-grained visual cues under cross-domain few-shot conditions. Consequently, a promising direction for future work is to improve the ability of the model to learn and exploit fine-grained visual representations in cross-domain few-shot learning scenarios.

CHAPTER 4

CONTINUAL TEST-TIME ADAPTATION VIA SOTA-DIT: SOURCE AND TARGET KNOWLEDGE DISENTANGLE TRANSFORMER

4.1 Introduction

Test-time adaptation (TTA) [19, 65, 77, 79, 83, 128, 170, 209, 211] addresses distribution shifts between the source training domain and the target test domain. Standard source-free TTA methods involve tuning a model with unlabelled test-time data to mitigate the domain shift problem. While effectively bridging the domain gap between the source and a single fixed target domain, they struggle with the constant target domain shifts at test time. Such constant distribution shifts of the test domain are common in real-world applications. For instance, an auto-driving model may encounter distribution shifts due to changes in weather, light conditions, and surrounding environments. These domain shifts hinder the model from achieving optimum performance.

Continual **T**est-**T**ime **A**daptation (CoTTA) [151, 175] is proposed to address the challenge of constant target domain distribution shifts. In CoTTA, an off-the-shelf pre-trained source model is provided to be tuned and tested with unlabelled target data from various target domains across different time steps. The objectives for better CoTTA are twofold: 1) prevent catastrophic forgetting: to preserve the source domain

knowledge over time without access to source data, and 2) extract high-quality target knowledge: to efficiently extract the target domain knowledge with only unlabelled test-time target domain data. Recent works mainly focus on preserving the source knowledge. For instance, tuning a more robust teacher model [117, 127, 175], training a stabilised model insensitive to the domain shifts [48, 128], or learning a group of prototypes based on the source model [17]. For target knowledge extracting, recent work facilitates an efficient sample selection strategy for negative and positive learning [129, 180]. However, recent works do not address both objectives simultaneously. They mainly focus on refining only one type of knowledge or extracting source and target knowledge in a mixed manner. In this thesis, we argue that this may harm knowledge preservation and extraction. Thus, we need to consider both types of knowledge and extract them separately.

To this end, we proposed a **So**urce and **Ta**rget knowledge **Di**stentangle **T**ransformer (SoTa-DiT) to explore knowledge disentangle for CoTTA utilising the prompting mechanism in a vision transformer (ViT) backbone, as illustrated in Figure 4.1. Inspired by the success of ProD in cross-domain few-shot image classification, we design a dual-prompt architecture for the transformer model to facilitate the disentanglement of source and target knowledge.

Specifically, in SoTa-DiT, we adopt the ViT [39] backbone because it enables the parallel use of multiple tiny, easy-to-plug-in vectors named visual prompts to extract knowledge separately, yielding a disentangled effect. The motivation of this prompt is designed similarly to ProD. Further, the standard student and teacher distillation architecture is adapted in SoTa-DiT for the ViT model as our baseline, following [175]. Based on that, the dual-prompt architecture is further designed to extract the source and target knowledge separately. More specifically, our SoTa-DiT has three key points:

First, SoTa-DiT employs a visual prompt named **S**ource **P**rompt (SP) to extract and preserve the source knowledge. SP is concatenated with the patched images to be fed into the transformer model. A group of loss functions are deliberately designed to tune SP and integrate the source knowledge into the model. We demonstrate that incorporating a contrastive loss between the augmented image feature embedded by the source model and the original image feature embedded by SP with the adapted model, along with a similarity loss between SP and the source model classification token, helps extract and preserve the source knowledge within SP. Additionally, implementing the symmetric cross-entropy loss [38] between the SP output and the model output while keeping the SP fixed during backpropagation effectively integrates the preserved source knowledge into the model.

Second, SoTa-DiT employs another visual prompt called **T**arget **P**rompt (TP) to extract target-domain knowledge. TP is also concatenated with patched images and fed into the transformer. A mask is applied to TP to prevent interactions between SP and TP. Following [38], we train TP with a contrastive loss conducted between the original and augmented image features. Additionally, a pseudo-label loss is conducted between the TP prediction and the combined TP+SP prediction to provide TP with the necessary label information.

Third, SoTa-DiT combines the source and target knowledge from SP and TP by simply averaging the SP and TP predictions after the softmax layers. Our evaluation demonstrates that this straightforward operation effectively integrates the source and target knowledge disentangled by visual prompts, resulting in a performance boost. By incorporating the three key points mentioned above, SoTa-DiT effectively preserves the source knowledge while extracting and integrating novel target domain knowledge. Consequently, SoTa-DiT adapts to the continually changing target domains effectively and significantly improves the classification accuracy on multiple datasets with different ViT backbones. For example, SoTa-DiT achieves average accuracy of 61.2% on ImageNet-C with ViT-B-16 backbone, +4.5% compared to the SOTA methods.

The contributions of SoTa-DiT can be summarized as follows:

• We proposed a novel **So**urce and **Ta**rget knowledge **Di**stentangle **T**ransformer (SoTa-DiT) method for continual test-time adaptation (CoTTA). SoTa-DiT disentangles the source and target knowledge through prompt learning.

• To prevent the catastrophic forgetting of the source knowledge, we designed a source prompt supervised by the source model using source contrastive and similarity loss. To effectively extract high-quality target knowledge, we designed a target prompt supervised by the target contrastive and cross-entropy loss. The two prompts extract two types of knowledge separately without direct interaction.

• We conduct comprehensive experiments to evaluate the effectiveness of source and target prompts in SoTa-DiT with different ViT backbones across various datasets. The results demonstrate the effectiveness of SoTa-DiT.

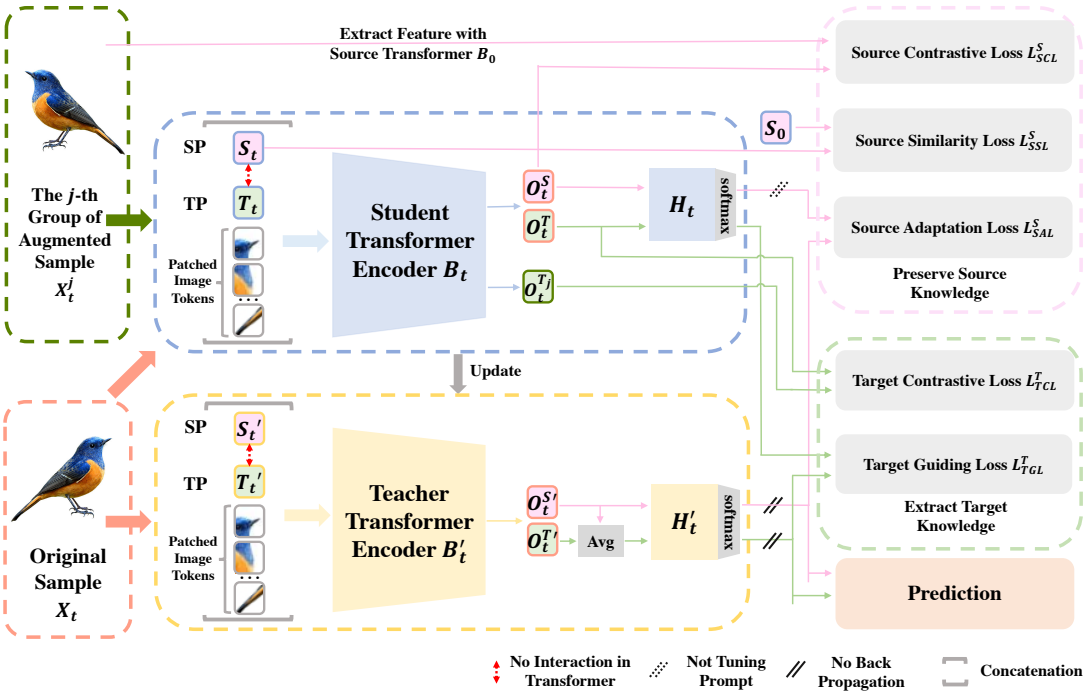


Figure 4.1: The overall architecture of SoTa-DiT involves two prompts, named source prompt (SP) and target prompt (TP), within a vision transformer (ViT) backbone using a teacher-student distillation architecture. At time step 0, both the teacher and student networks are initialised with the source model. Meanwhile, the prompts from both networks copy the source classification token. At time step t, we feed the original and augmented images to get prompt outputs O_t. The augmented images are also fed into the source model to get the source token output. These outputs are then utilized to tune 1) SP by calculating a source contrastive loss $\mathscr{L}_{SCL}^S$ and a source similarity loss $\mathscr{L}_{SSL}^S$ to preserve source knowledge, and 2) TP by calculating a target contrastive loss $\mathscr{L}_{TCL}^T$ to extract target knowledge and a target guiding loss $\mathscr{L}_{TGL}^T$ to learn label information. Furthermore, a source adaptation loss $\mathscr{L}_{SAL}^S$ is calculated to adapt the preserved source knowledge to the model. Finally, we average the predictions of SP and TP from the teacher model to obtain the final prediction.

4.2 Methodology

4.2.1 Problem Formulation

In this section, we focus on the image classification task under the source-free CoTTA setting. We start with a source model $f_{\theta_s}(x)$ pre-trained on the source domain Φ_S with data $(\mathscr{X}^S, \mathscr{Y}^S)$. Under the source-free setting, $(\mathscr{X}^S, \mathscr{Y}^S)$ is unavailable for fine-tuning. Given $f_{\theta_s}(x)$, our objective is to achieve a higher image classification accuracy when

testing on continually changing target domains. At test time, we have unlabelled target data batches $\mathscr{X}^T = (X_1, X_2, ...)$ from test domains different from Φ_S. These target data are presented to the model sequentially, following a time-step sequence. At a time step t, f_{θ_t} classify target data batch X_t. Meanwhile, X_t is used to tune $f_{\theta_t}(x)$ for the next time step $t+1$. Note that the distribution of X_t is constantly changing, and the model is evaluated in an online manner as it adapts dynamically to each new data batch.

4.2.2 Overall Architecture

The overall architecture of **So**urce and **Ta**rget knowledge **Di**stentangle **T**ransformer (SoTa-DiT) is illustrated in Figure 4.1. In SoTa-DiT, we utilize a student model, denoted as $f_\theta = (B, \mathbf{S}, \mathbf{T}, H)$ and a teacher model, denoted as $f_{\theta'} = (B', \mathbf{S}', \mathbf{T}', H')$. Each model consists of a transformer, denoted as B, a **S**ource **P**rompt (SP), denoted as $\mathbf{S}$, a **T**arget **P**rompt (TP), denoted as $\mathbf{T}$, and a classification head, denoted as H. At the time step 0, f_{θ_0} and $f_{\theta_0'}$ are both initialized with the source model f_{θ_s}. The visual prompts, specifically S_0, T_0, $S_0{}'$ and $T_0{}'$, are initialized with the classification token from the source model f_{θ_s}.

At time step t, we present a test data batch $X_t = \{x_1, x_2, ...x_{N_t}\}$, comprising N_t images from C categories. The data batch is sampled from an unknown domain with a distribution different from the source dataset. Each image x_i is first processed into patched image tokens $G_i \in \mathbb{R}^{K \times D}$, where K represents the number of tokens and D denotes the dispatch embedding dimension. For $\mathbf{S}$ and $\mathbf{T}$, we set $\mathbf{S}, \mathbf{T} \in \mathbb{R}^{1 \times D}$ to ensure their compatibility with the embedding dimension of the patched image tokens. Then, a mask is applied to $\mathbf{S}$ and $\mathbf{T}$ to prevent the direct interaction between them within B_t and $B_t{}'$. Finally, we concatenate $\mathbf{S}$ and $\mathbf{T}$ to G and feed them into f_{θ_t} and $f_{\theta_t'}$, following:

$$\begin{aligned} (O_i^S, O_i^T, E_i) &= B_t([\mathbf{S}_t, \mathbf{T}_t, G_i]) \\ (O_i^{S'}, O_i^{T'}, E_i{}') &= B_t{}'([\mathbf{S}_t{}', \mathbf{T}_t{}', G_i]) \end{aligned} \tag{4.1}$$

where [ ], O, and E represent concatenation operation, prompt output, and patched image embedding after transformer blocks, respectively.

Following the discovery that the feature tokens are not suitable for classification in the section above (ProD), E is discarded and only the prompt output O are used for subsequent operations. We denote the output of $\mathbf{S}$ and $\mathbf{T}$ for X_t^T from f_{θ_t} and $f_{\theta_t'}$ as $\mathbf{O}_t^S = (O_1^S, O_2^S, ...O_{N_t}^S)$, $\mathbf{O}_t^T = (O_1^T, O_2^T, ...O_{N_t}^T)$, $\mathbf{O}_t^{S'} = (O_1^{S'}, O_2^{S'}, ...O_{N_t}^{S}{}')$ and $\mathbf{O}_t^{T'} = (O_1^{T'}, O_2^{T'}, ...O_{N_t}^{T}{}')$, respectively. Then, $\mathbf{O}_t^{S'}$ and $\mathbf{O}_t^{T'}$ are fed into the teacher classification head $H_t{}'$ and a softmax layer to get the prediction $\mathbf{P}_t^{S'} = (P_1^{S'}, P_2^{S'}, ...P_{N_t}^{S}{}')$ and

$\mathbf{P}_t^{T\prime} = (P_1^{T\prime}, P_2^{T\prime}, ...P_{N_t}^{T\prime})$. Subsequently, $\mathbf{P}_t^{S\prime}$ and $\mathbf{P}_t^{T\prime}$ are averaged to obtain the final inference prediction $\mathbf{P}_t$ for X_t and the classification accuracy is calculated.

Meanwhile, the student network f_{θ_t} is tuned for the next time step $t+1$ using the prompt output $\mathbf{O}$ from the time step t. The specific procedure of loss calculation for both prompts is explained in the following subsections. During the back-propagating procedure, B_t and H_t are updated with learning rate δ, while $\mathbf{S}$ and $\mathbf{T}$ are updated with learning rate $\mu \times \delta$. Here, μ is introduced to provide prompts with an extra learning rate, enhancing their ability to extract the knowledge. After updating the student network $f_{\theta_{t+1}}$, we update teacher network $f_{\theta_{t'}}$ using the exponential moving averages (EMA):

$$f_{\theta_{t+1}\prime} = \gamma f_{\theta_t'} + (1-\gamma) f_{\theta_{t+1}}, \tag{4.2}$$

where hyper-parameter γ controls the updating speed.

4.2.3 Source Prompt for Source Knowledge Preservation

SoTa-DiT utilises source prompt, denoted as $\mathbf{S}$, to preserve the source knowledge stored within the source model and adapt it to other parts of the model. $\mathbf{S}$ is trained with a source contrastive loss $\mathscr{L}_{SCL}^S$ and a source similarity loss $\mathscr{L}_{SSL}^S$ to preserve the source knowledge. Then, we apply a symmetric cross-entropy loss called source adaptation loss $\mathscr{L}_{SAL}^S$ to adapt the source knowledge to the other parts of the model. The loss calculation details are introduced as follows.

4.2.3.1 Source Contrastive Loss

At the time step t, SoTa-DiT obtains the SP output of the original test image batch from the student network, denoted as $\mathbf{O}_t^S = (O_1^S, O_2^S, ...O_{N_t}^S)$. Meanwhile, each image from the test data batch X_t is randomly augmented M times, generating M groups of augmented data $X_t^1, X_t^2, ...X_t^M$. For each group of augmented data X_t^j, we have $X_t^j = (x_1^j, x_2^j, ...x_{N_t}^j)$. Then, the augmented data are processed into dispatched image tokens and fed into the source transformer B_0 with $\mathbf{S}_t$ being concatenated, following Eq.4.1. Finally, we obtain the output of the augmented data X_t^j from the source model, denoted as $\mathbf{O}_t^{S_j} = (O_1^{S_j}, O_2^{S_j}, ...O_{N_t}^{S_j})$. Finally, the source contrastive loss $\mathscr{L}_{SCL}^S$ is calculated as:

$$\mathrm{L}_{SCL}^S = -\frac{1}{M}\sum_{j=1}^{M}\sum_{i=1}^{N_t}\frac{\exp\left(\mathrm{sim}(O_i^S, O_i^{S_j})/\tau_s\right)}{\sum_{k=1}^{N_t}\exp\left(\mathrm{sim}(O_i^S, O_k^{S_j})/\tau_s\right)}, \tag{4.3}$$

where $\exp(x) = e^x$, and $\text{sim}(z_i, z_j) = \frac{z_i \cdot z_j}{\|z_i\| \times \|z_j\|}$.

With $\mathscr{L}^S_{SCL}$, **S** extracts and preserves the source knowledge from the source model. However, **S** can still become unstable as the time step increases. To further stabilise the tuning process of **S**, we apply a simple but effective loss of source similarity to restrict **S**.

4.2.3.2 Source Similarity Loss

To stabilise the tuning procedure of **S**, SoTa-DiT constrains the source prompt S_t at each time step t with the source prompt S_0 from the initial time step using a similarity loss. Notice that the initial source prompt is the same as the classification token from the source model. The similarity loss $\mathscr{L}^S_{SSL}$ is calculated as followed:

$$\mathscr{L}^S_{SSL} = 1 - |\frac{S_t \cdot S_0}{\|S_t\| \times \|S_0\|}|, \tag{4.4}$$

.

The source similarity loss ensures that the shape of $\mathbf{S}_t$ remains similar to the source model classification token. This restriction stabilises the tuning of **S**. Further, we allow $\mathbf{S}_t$ to be reasonably elastic by adding a weight parameter.

4.2.3.3 Source Adaptation Loss

With $\mathscr{L}^S_{SCL}$ and $\mathscr{L}^S_{SSL}$, the source prompt **S** extracts the source knowledge from the source model and preserves it effectively. To further adapt the preserved source knowledge to other parts of the model, we apply a source adaptation loss, denoted as $\mathscr{L}^S_{SAL}$. During the backpropagation, $\mathscr{L}^S_{SAL}$ updates all model parameters except for the source prompt **S**. This process ensures that **S** adapts the source knowledge to other network parts without losing the preserved knowledge.

At the time step t, SoTa-DiT first obtains the SP output of the image batch from both the student and teacher networks, denoted as $\mathbf{O}^S_t = (O^S_1, O^S_2, ...O^S_{N_t})$ and $\mathbf{O}^{S'}_t = (O^{S'}_1, O^{S'}_2, ...O^{S\ '}_{N_t})$, respectively. We then fed $\mathbf{O}^S_t$ and $\mathbf{O}^{S'}_t$ into the student classification H_t and teacher classification head $H_t{}'$ along with softmax layers. This process generates the SP predictions of the student and teacher models for the C image categories, denoted as $\mathbf{P}^S_t = (P^S_1, P^S_2, ...P^S_{N_t})$ and $\mathbf{P}^{S'}_t = (P^{S'}_1, P^{S'}_2, ...P^{S\ '}_{N_t})$, respectively. Here, $P^S_i = (p^S_1, p^S_2, ...p^S_C)$, and $P^{S'}_i = (p^{S'}_1, p^{S'}_2, ...p^{S'}_C)$, where p^S_i and $p^{S'}_i$ represent the possibilities that a test sample belongs to each category. Finally, the source adaptation loss $\mathscr{L}^S_{SAL}$ is calculated between $\mathbf{P}^S_t$ and $\mathbf{P}^{S'}_t$, given by:

$$\mathscr{L}^S_{SAL} = -\frac{1}{N_t C}\sum_{i=1}^{N_t}(\frac{1}{2}P_i^{S'}\cdot\log P_i^S + \frac{1}{2}P_i^S\cdot\log P_i^{S'}), \tag{4.5}$$

.

The overall loss for **S** is formulated as:

$$\mathscr{L}^S = \mathscr{L}^S_{SAL} + \alpha\mathscr{L}^S_{SCL} + \eta\mathscr{L}^S_{SSL}, \tag{4.6}$$

where hyper-parameter α and η controls the loss weights.

4.2.4 Target Prompt for Target Knowledge Extraction

SoTa-DiT utilises the target prompt, denoted as **T**, to extract target knowledge from the unlabelled test data efficiently. Specifically, **T** is trained with a target contrastive loss $\mathscr{L}^T_{TCL}$ and a target guiding loss $\mathscr{L}^T_{TGL}$, as described in the following two subsections.

4.2.4.1 Target Contrastive Loss

SoTa-DiT first obtains the TP output of the original image from the student network, denoted as $\mathbf{O}^T_t = (O^T_1, O^T_2, ...O^T_{N_t})$. Similar to the source contrastive loss calculation, we randomly augment the test data batch X_t, generating M groups of augmented data $X^1_t, X^2_t, ...X^M_t$. The augmented data are then processed and fed into the student transformer B_t with $\mathbf{T}_t$ being concatenated, following Eq.4.1.

Then, SoTa-DiT obtains the TP output of the augmented data X^j_t, denoted as $\mathbf{O}^{T_j}_t = (O^{T_j}_1, O^{T_j}_2, ...O^{T_j}_{N_t})$. Finally, the target contrastive loss $\mathscr{L}^T_{TCL}$ is calculated as:

$$\mathrm{L}^T_{TCL} = -\frac{1}{M}\sum_{j=1}^{M}\sum_{i=1}^{N_t}\frac{\exp\left(\mathrm{sim}(O^T_i, O^{T_j}_i)/\tau_t\right)}{\sum_{k=1}^{N_t}\exp\left(\mathrm{sim}(O^T_i, O^{T_j}_k)/\tau_t\right)}, \tag{4.7}$$

.

With the contrastive loss $\mathscr{L}^T_{TCL}$, $\mathbf{T}_t$ extracts target knowledge by clustering the test images based on their similarities. However, without further label information, $\mathbf{T}_t$ may struggle to predict the actual image label. To address this, we design a guiding loss to provide $\mathbf{T}_t$ with label information.

4.2.4.2 Target Guiding Loss

SoTa-DiT uses both $\mathbf{S}_t$ and $\mathbf{T}_t$ to generate a label prediction to guild $\mathbf{T}_t$. The teacher TP and SP outputs, denoted as $\mathbf{O}_t^{T'}$ and $\mathbf{O}_t^{S'}$, are fed into the teacher classification head $H_t{}'$ to generate the label prediction $P_i^{T+S'} = (p_1^{T+S'}, p_2^{T+S'}, ...p_C^{T+S'})$, given by:

$$P_i^{T+S'} = \text{softmax}\big(H_t{}'(\frac{\mathbf{O}_t^{T'} + \mathbf{O}_t^{S'}}{2})\big). \tag{4.8}$$

Then, SoTa-DiT feeds the TP output $\mathbf{O}_t^T = (O_1^T, O_2^T, ...O_{N_t}^T)$ into the student classification head H_t and a softmax layer to obtain the TP prediction $\mathbf{P}_t^T = (P_1^T, P_2^T, ...P_{N_t}^T)$, with $P_i^T = (p_1^T, p_2^T, ...p_C^T)$. Finally, the target guiding loss $\mathscr{L}_{TGL}^T$ is calculated using $\mathbf{P}_t^T$ and $P_i^{T+S'}$, given by:

$$\mathscr{L}_{TGL}^T = -\frac{1}{N_t C}\sum_{i=1}^{N_t}(\frac{1}{2}P_i^{T+S'} \cdot \log P_i^T + \frac{1}{2}P_i^T \cdot \log P_i^{T+S'}). \tag{4.9}$$

The target guiding loss enables **T** to produce more accurate label predictions with target knowledge. The overall loss for the target prompt **T** is formulated as:

$$\mathscr{L}^T = \mathscr{L}_{TCL}^T + \beta \mathscr{L}_{TGL}^T, \tag{4.10}$$

where hyper-parameter β controls the loss weights.

4.2.4.3 Overall Loss

Finally, the overall loss function $\mathscr{L}$ is formulated as:

$$\mathscr{L} = \delta L^S + L^T, \tag{4.11}$$

where hyper-parameter δ controls the overall loss weights.

4.3 Experiments

4.3.1 Settings

Datasets. SoTa-DiT is evaluated on three CoTTA datasets: ImageNet-C [69], ImageNet-R [68], and ImageNet-D109 [129]. ImageNet-C includes 15 types of image corruption with

Table 4.1: The averaged image classification accuracy (%) of different methods with various transformer backbones on multiple datasets. IN-C, IN-R, and IN-D109 represent ImageNet-C, ImageNet-R, and ImageNet-D109, respectively. 'Source' represents the backbone model without adaptation. SoTa-DiT outperforms other counterpart methods with clear margins across three different datasets with three commonly used ViT backbones on CoTTA.

Dataset	**Backbone**	SOURCE	MEMO [211]	DDA [49]	TENT [170]	ETEA [128]	COTTA [175]	RMT [38]	SAR [129]	ROID [117]	SoTa-DiT
IN-C	ViT-S-16	28.2	39.8	41.0	43.1	43.9	43.8	44.1	44.5	45.1	**48.1**
	ViT-B-16	42.8	50.9	52.2	54.0	56.0	54.6	54.7	56.2	56.7	**61.2**
	ViT-L-16	46.2	52.9	51.6	59.3	61.2	59.9	62.1	62.4	63.9	**70.4**
IN-R	ViT-S-16	32.4	33.1	32.0	36.7	45.0	41.2	46.3	45.1	47.8	**51.2**
	ViT-B-16	44.0	45.4	45.6	46.7	51.0	30.4	31.2	51.4	55.8	**60.2**
	ViT-L-16	51.2	50.2	52.3	53.6	56.7	30.2	34.8	55.9	62.8	**69.4**
IN-D109	ViT-S-16	39.8	41.4	42.6	9.6	44.9	20.0	21.2	38.7	47.8	**50.2**
	ViT-B-16	46.4	40.2	47.0	16.0	52.6	26.6	25.8	42.6	55.0	**58.2**
	ViT-L-16	57.4	39.8	37.6	17.2	61.9	25.1	30.2	53.3	62.0	**68.2**

Table 4.2: The image classification accuracy (%) of different methods using Vit-B-16 backbone on the ImageNet-C dataset for 15 different types of corruption. SoTa-DiT achieves SOTA performance, exceeding other methods by clear margins on 10 of 15 corruptions.

Method	gaussian	shot	impulse	defocus	glass	motion	zoom	snow	frost	fog	bright	contrast	elsatic	pixelate	jpeg	**Avg.**
SOURCE	46.2	30.5	33.3	31.2	27.1	44.3	30.3	53.3	48.2	45.3	75.2	8.9	44.0	60.8	62.7	42.8
MEMO	53.4	31.2	35.4	52.2	44.7	50.6	42.8	37.5	48.8	60.1	68.2	63.2	51	62.4	62.1	50.9
DDA	52.2	30.9	38.4	50.0	45.2	50.1	46.1	38.8	49.2	66.1	70.2	65.4	56.0	62.3	61.5	52.2
TENT	51.3	34.9	39.9	56.2	43.2	54.8	50.1	40.1	52.9	60.0	74.5	68.1	53.4	62.7	67.7	54.0
ETEA	51.2	35.5	40.2	**58.7**	44.5	56.1	49.2	42.8	53.8	68.9	**78.4**	**68.4**	58.0	65.5	68.2	56.0
COTTA	52.7	34.2	40.8	53.4	45.2	55.1	49.8	40.2	50.6	**69.1**	76.6	66.4	55.2	64.1	65.1	54.6
RMT	55.4	63.9	56.6	50.6	54.5	57.1	44.7	50.6	51.8	47.8	76.2	31.0	56.5	65.5	58.4	54.7
SAR	51.8	36.4	41.5	53.7	46.7	57.5	**52.8**	50.1	50.7	68.1	74.6	65.7	57.9	68.9	65.9	56.2
ROID (sota)	54.8	36.2	58.2	55.4	46.2	57.1	50.2	42.9	57.2	62.3	77.6	67.2	53.4	64.2	67.9	56.7
SoTa-DiT (ours)	**63.8**	**66.5**	**65.0**	57.6	**59.9**	**61.0**	51.0	**61.8**	**62.6**	53.8	76.5	33.9	**65.7**	**70.5**	**69.0**	**61.2**

5 severity levels. We conduct the evaluations under the level 5 corruption. ImageNet-R includes 30000 images with different renditions from 200 categories. ImageNet-D109, a subset of ImageNet-D [136, 147], contains 109 classes with six types of domain shifts. SoTa-DiT is compared with the other methods on these three datasets, and ablation studies are conducted on the ImageNet-C dataset.

Implementation Details. We evaluate SoTa-DiT with three different ViT backbones: Vit-B-16, ViT-S-16, and ViT-L-16 [39]. The off-the-shelf source model are pretrained on ImageNet [34], following [24, 155, 164, 207, 221]. The ablation studies are

conducted on the Vit-B-16 backbone. During the test-time tuning, the image batches are provided to the source network in an online manner, following [175]. The network predicts the image category and adapts to the test domain, which constantly changes over time. We set the learning rate to 0.001 and the tuning batch size to 8. γ (Eq.4.2) and the training step are set to 0.999 and 1. μ, τ_s (Eq.4.3), and τ_t (Eq.4.7) are set to 20, 0.05 and 0.05 by default. δ (Eq.4.11) is set to 1.1. α, η (Eq.4.6) and β (Eq.4.10) are set to 1.0, 0.9 and 1.0 by default. M (Eq.4.7) is set to 2 by default.

4.3.2 Effectiveness of SoTa-DiT

We compare SoTa-DiT with the baseline and the SOTA methods in Table 4.1. We draw three observations.

First, SoTa-DiT outperforms the SOTA method across multiple datasets with three most commonly used ViT backbones: ViT-S-16, ViT-B-16, and ViT-L-16. Specifically, on the ViT-B-16 backbone, SoTa-DiT outperforms the SOTA methods on ImageNet-C, ImageNet-R, and ImageNet-D109 by +4.5%, +4.4%, and +3.2%, respectively. The result proves that SoTa-DiT as a whole is superior for the CoTTA task.

Second, SoTa-DiT demonstrates more substantial performance advantages with larger ViT backbones. For instance, on ImageNet-C, SoTa-DiT with ViT-S-16, ViT-B-16, and ViT-L-16 outperforms the SOTA methods by +3.0%, +4.5%, and +6.5%, respectively. We infer that the reason for this is that the larger ViT model encapsulates more general knowledge. Thus, the SP absorbs and preserves more domain-general source knowledge that is beneficial for the CoTTA task.

Third, SoTa-DiT consistently outperforms other methods across various types of corruption on the ImageNet-C dataset. As detailed in Table 4.2, SoTa-DiT outperforms other methods on 10 out of 15 types of corruption. This result proves that the SoTa-DiT is general and suitable for different corruptions for the test-time adaptation task.

4.3.3 Ablation Studies of Key Components

4.3.3.1 Source and Target Prompts

We examine the effectiveness of source prompt (SP), target prompt (TP), and their associated loss functions on the ImageNet-C dataset, detailed in Table 4.3. The 'Baseline' refers to the ViT adapted with the method from [175]. In the 'SoTa-DiT*', we let $\mathscr{L}^S_{SAL}$ tune SP during the backpropagation. We draw four observations based on the results:

Table 4.3: Evaluation of the key components, including the Source Prompt (SP), Target Prompt (TP), and loss functions, on ImageNet-C. We list the averaged classification accuracy (%). The result proves that both SP and TP, along with their loss functions, are effective and are beneficial for CoTTA.

Method	SP	TP	$\mathscr{L}^S_{SCL}$	$\mathscr{L}^S_{SSL}$	$\mathscr{L}^S_{SAL}$	$\mathscr{L}^T_{TCL}$	$\mathscr{L}^T_{TGL}$	Average Acc.
Baseline (CoTTA)	✗	✗	✗	✗	✗	✗	✗	54.6
SP only	✓	✗	✓	✓	✓	✗	✗	55.1
SP only-$\mathscr{L}^S_{SCL}$	✓	✗	✗	✓	✓	✗	✗	54.8
SP only-$\mathscr{L}^S_{SSL}$	✓	✗	✓	✗	✓	✗	✗	54.9
SP only-$\mathscr{L}^S_{SAL}$	✓	✗	✓	✓	✗	✗	✗	54.7
TP only	✗	✓	✗	✗	✗	✓	✓	58.9
TP only-$\mathscr{L}^T_{TCL}$	✗	✓	✗	✗	✗	✗	✓	54.9
TP only-$\mathscr{L}^T_{TGL}$	✗	✓	✗	✗	✗	✓	✗	50.9
TP only+$\mathscr{L}^S_{SCL}$	✗	✓	✓	✗	✗	✓	✓	59.2
TP only+ $\mathscr{L}^S_{SSL}$	✗	✓	✗	✓	✗	✓	✓	59.1
SoTa-DiT*	✓	✓	✓	✓	✓	✗	✗	60.4
SoTa-DiT	✓	✓	✓	✓	✓	✓	✓	**61.2**

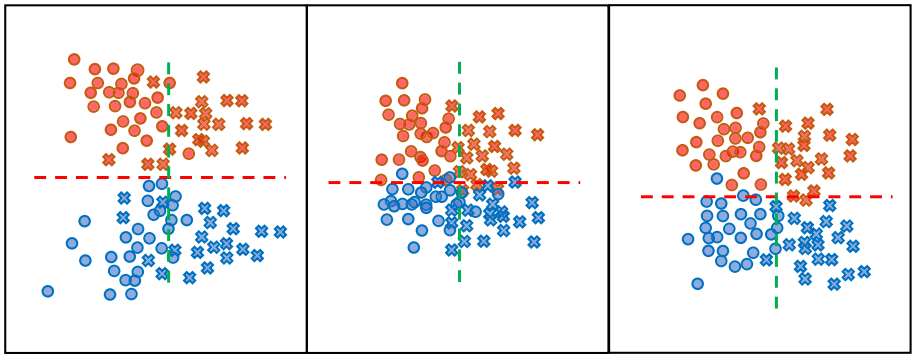
First, training with only SP or TP improves performance. Adding SP to the baseline increases the average accuracy by +0.5%, demonstrating that learning the source knowledge with SP helps to prevent forgetting and benefits the CoTTA task. Similarly, adding TP to the baseline increases the average accuracy by +4.3%. This result proves that extracting target knowledge with TP also benefits the CoTTA task.

Second, comparing 'TP only+$\mathscr{L}^S_{SCL}$' with 'TP only,' we see that directly adding a source contrastive loss to the TP benefits CoTTA, with the accuracy increased by +0.3%. This result shows that using $\mathscr{L}^S_{SCL}$ can effectively preserve the source knowledge within different prompts.

Third, comparing 'SoTa-DiT*' with 'SoTa-DiT' reveals that letting $\mathscr{L}^S_{SAL}$ tune the SP is sub-optimal, with a performance decrease of -0.8%. We infer that the preserved source knowledge within SP is harmed by $\mathscr{L}^S_{SAL}$. It is because when SP is tuned with $\mathscr{L}^S_{SAL}$, the target knowledge is absorbed by SP, which contradicts the purpose of purposing SP.

Finally, jointly using SP and TP brings a significant performance boost compared to using SP (+6.1%) or TP alone (+2.3%). The result proves that extracting source and target knowledge in a disentangling manner is beneficial to the CoTTA task.

Additionally, we visualised the feature embedding of samples from two image categories with two different types of corruption in ImageNet-C, as shown in Figure 4.2.



(a) TP Only **(b) SP Only** **(c) SoTa-DiT**

Figure 4.2: Visualisation of image embeddings from two different categories, represented by crosses and circles. The images are from two different domains, colored in red and blue. The domain and class boundaries are the red and green lines, respectively. Samples are selected from the ImageNet-C dataset. TP embedding contains more target knowledge and shows more precise domain boundaries (a), while SP embedding contains more source knowledge and shows more precise class boundaries (b). Jointly using them yields more precise domain boundaries with refined class boundaries.

The figure illustrates that using TP alone separates the image features from each other based on domain and class, which aids classification. However, several samples are misclassified and appear distant from their true class centres, possibly due to overfitting the extreme cases with TP alone and forgetting source knowledge. Conversely, using SP alone does not push the domain and class centres as far as using TP alone. However, the sample misclassification issue is slightly alleviated. When SP and TP are used together in SoTa-DiT, the final embeddings show clearer boundaries between categories and domains, with fewer samples misplaced in other clusters. The observation result proves that the two prompts disentangle the source and target knowledge, and the two types of knowledge have different characteristics. By using them together, SoTa-DiT combines the advantages of both and generates better feature representations for the CoTTA task.

4.3.3.2 Knowledge Combining

We investigate different ways of combining source and target knowledge, as detailed in Table 4.4. The table denotes different methods: 1) 'SP only': Utilize SP predictions alone. 2) TP only: Utilize TP predictions alone. 3) 'Avg. before H': Average SP and TP outputs before the classification head H. 4) 'SoTa-DiT': Average the SP and TP predictions after

Table 4.4: Evaluation of different ways to combine source and target knowledge on ImageNet-C. We list the classification accuracy (%). The result shows that combining the two features after the classification head is optimum. The reason is that SP and TP features are complementary and after classification head with softmax, the strengths of both features are reinforced.

Method	Average Acc.
Baseline (CoTTA)	54.6
SP only	56.2
TP only	59.9
Avg. before H	60.4
SoTa-DiT	**61.2**

the softmax layer. We draw three observations:

Firstly, both SP and TP enhance the prediction accuracy after disentangling the source and target knowledge compared to the baseline. For instance, comparing 'SP only' from Table 4.4 and 'SP only' from Table 4.3, we observe an improvement of +1.1%. Similarly, the 'TP only' also shows an increase in accuracy by +1.0%. This proves that disentangled knowledge is of a higher quality.

Secondly, comparing 'Avg. before H' and 'SoTa-DiT', we observe that averaging the predictions after the softmax layers brings higher accuracy (+0.8%). We infer that the target and source knowledge may provide better predictions for different samples. After the softmax layer, the advantages of both types of knowledge are further amplified.

Third, comparing 'SoTa-DiT' with 'SP only' and 'TP only', we find that jointly utilizing the disentangled source and target knowledge brings further improvement, with accuracy increasing by +5.0% and +1.3%. This highlights the advantages of combining knowledge for CoTTA.

4.3.4 Ablation Studies of Hyper-parameters

4.3.4.1 Source Contrastive Loss Weight α

We evaluate the influence of the source contrastive loss weight α in Eq.4.6 for SP. The results are illustrated in Figure 4.3 with a blue line. We draw the following observation:

As α increases, the accuracy increases and then decreases. The reason is two-fold. First, when α is below the optimum value of 1.0, the weight of source contrastive loss is low, and the ability of SP to extract and preserve source knowledge decreases. The model cannot preserve the knowledge, leading to catastrophic forgetting that harms

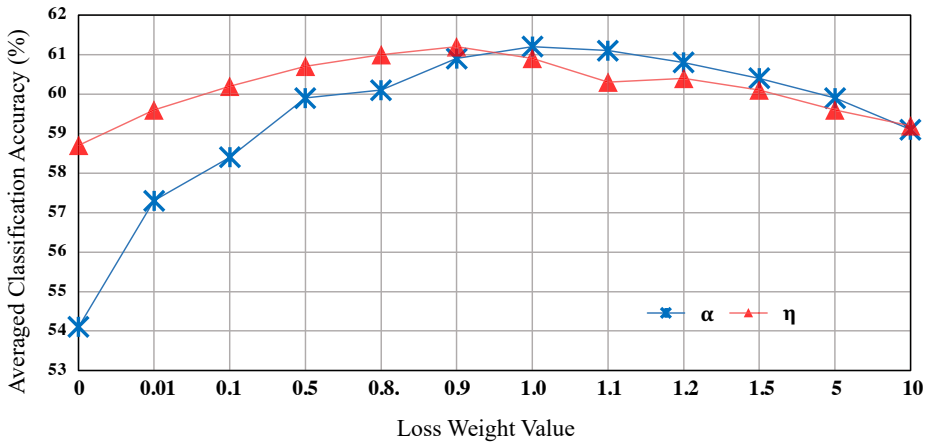


Figure 4.3: Averaged classification accuracy of SoTa-DiT with different source contrastive loss and source similarity loss weight, named 'α' and 'η', for the source prompt on ImageNet-C. The result shows that $\alpha = 1$ and $\eta = 0.9$ yield the highest classification accuracy. The effects of these two parameters share a similar trend since they are proportional to the source knowledge preservation strength.

CoTTA. Conversely, as α exceeds 1.0, the emphasis on the source contrastive loss increases, potentially overshadowing the cross-entropy loss. Consequently, the preserved source knowledge is not adapted to the other parts of the model effectively, leading to catastrophic forgetting in other parts of the model other than the SP.

4.3.4.2 Source Similarity Loss Weight η

We evaluate the influence of the source similarity loss weight η in Eq.4.6 for SP. The results are illustrated in Figure 4.3 with a red line. We draw the following observation:

Similar to the effect observed with α, as η increases, the accuracy increases and then decreases. We infer the reason as two-fold. When η is below 0.9, similarity loss cannot effectively constrain the shape of SP, leading to catastrophic forgetting. When η exceeds 0.9, similarity loss dominates. As a result, SP loses its elasticity and cannot adapt the preserved source knowledge to the target domain.

4.3.4.3 Target Guiding Loss Weight β

We evaluate the influence of the target guiding loss weight β in Eq.4.10 for TP. The results are depicted in Figure 4.4 with a green line. We draw the following observation:

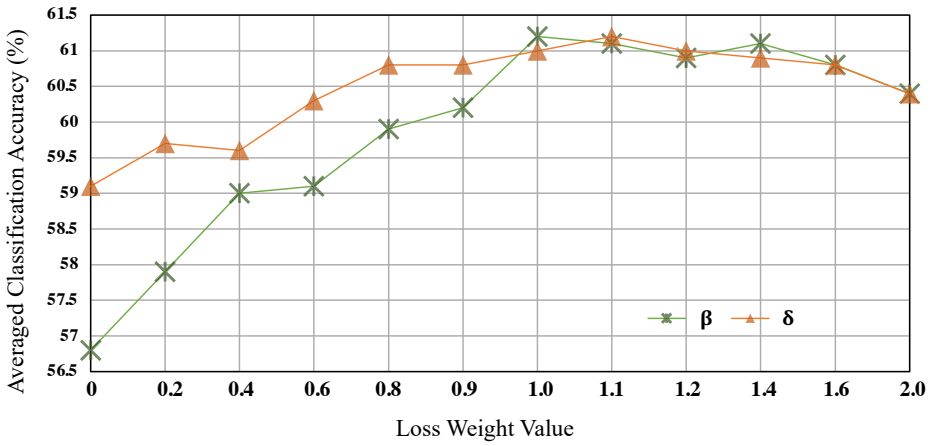


Figure 4.4: Averaged classification accuracy of SoTa-DiT with different target guiding loss weight 'β' for the target prompt and overall loss trade-off weight 'δ' on ImageNet-C. The result shows that $\beta = 1$ and $\delta = 1.1$ yield the highest classification accuracy.

As β increases, the accuracy increases significantly and then decreases. We infer that two factors contribute to this. Firstly, the guiding loss is inactive when β is less than 1.0. Consequently, TP is unable to learn label information effectively. As the target contrastive loss only clusters the samples without providing label information, TP may match sample clusters with random labels, leading to a severe mismatch between samples and their labels. When β exceeds 1.0, the contrastive loss is partially deactivated, hindering the effective extraction of target knowledge. As a result, TP cannot effectively learn the target domain knowledge of higher quality, leading to a decrease in classification accuracy.

4.3.4.4 Overall Loss Trading-off Weight δ

We evaluate the influence of the overall loss trade-off weight δ in Eq.4.10. The results are depicted in Figure 4.4 with an orange line. We draw the observation:

As δ increases, the accuracy increases gradually and then decreases. We infer that two factors contribute to this. When δ is smaller than 1.1, the weight of the source prompt loss is low. SP is not properly tuned and thus cannot preserve enough source knowledge. Moreover, the preserved source knowledge is not effectively adapted to other parts of the model. Hence, the model is subject to catastrophic forgetting and the classification accuracy decreases. Conversely, when δ exceeds 1.1, the weight of the target prompt loss is relatively low. TP is not tuned properly and thus cannot extract enough target knowledge. Thus, the model cannot effectively adapt to the incoming novel domains.

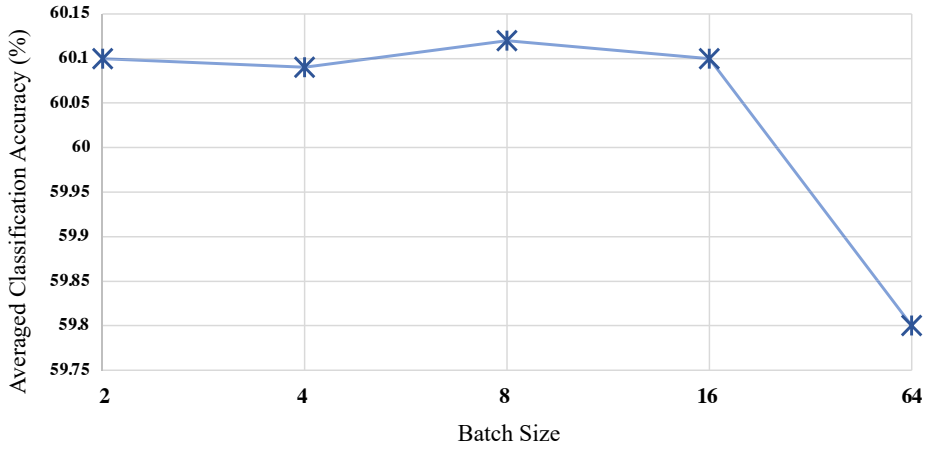


Figure 4.5: Averaged image classification accuracy of SoTa-DiT with different batch sizes on ImageNet-C. The result shows that batch size being set as 8 yields the highest classification accuracy.

4.3.4.5 Batch Size

We investigate the influence of batch size on accuracy. The results are depicted in Figure 4.5. We draw the following observation:

As batch size increases, accuracy initially remains stable and decreases. We infer that the reason behind this phenomenon is that the large batch size influences TP's target contrastive loss. When the batch size is too large, contrastive learning for TP becomes more challenging as the total number of test samples is limited, with minor inter-batch diversity. Consequently, the target contrastive loss cannot effectively supervise TP, leading to a decline in performance.

4.3.4.6 Prompt Extra Learning Rate μ

We evaluate the influence of prompt extra learning rate μ. The results are depicted in Figure 4.6. We draw the following observation.

As the extra learning rate μ increases, the accuracy increases sharply and decreases slightly. The reason behind this is twofold. On one hand, when the μ is lower than 1, TP and SP learn slower than other parts of the network. As a result, they fail to extract the necessary source and target knowledge, resulting in a performance decline. On the other hand, as μ becomes too large, the learning rate of TP and SP is large. As a result, they may miss the actual optimum points and be less robust to the perturbation from the test samples.

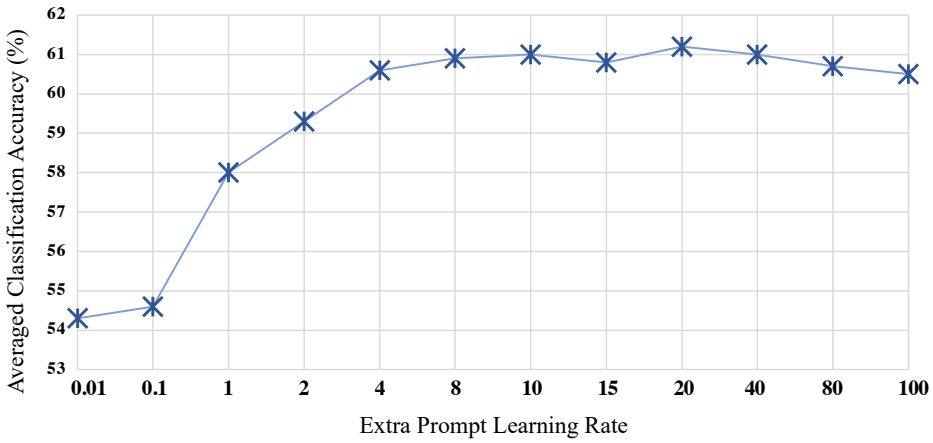


Figure 4.6: Averaged image classification accuracy of SoTa-DiT with different prompt extra learning rates μ on ImageNet-C. The result shows that $\mu = 20$ yields the highest classification accuracy.

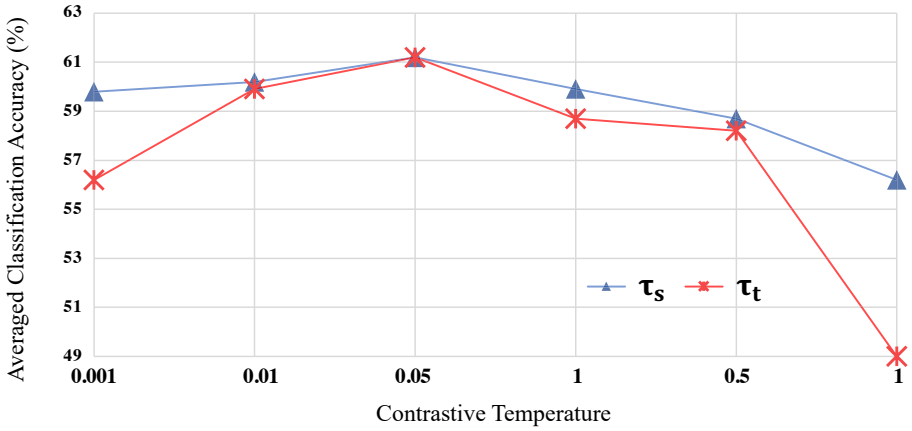


Figure 4.7: Averaged classification accuracy of SoTa-DiT with different contrastive temperatures τ_s and τ_t on ImageNet-C. The result shows that $\tau_s = 0.05$ and $\tau_t = 0.05$ yield the highest classification accuracy.

4.3.4.7 Contrastive Temperature τ

We evaluate the influence of contrastive temperature τ. The results are depicted in Figure 4.7. The average accuracy under different contrastive temperatures for source contrastive loss (τ_s)and target contrastive loss (τ_t) is depicted with blue and red lines, respectively. Based on the result, we draw the following observations.

First, as τ_s and τ_t increase, the accuracy initially increases, then decreases. The

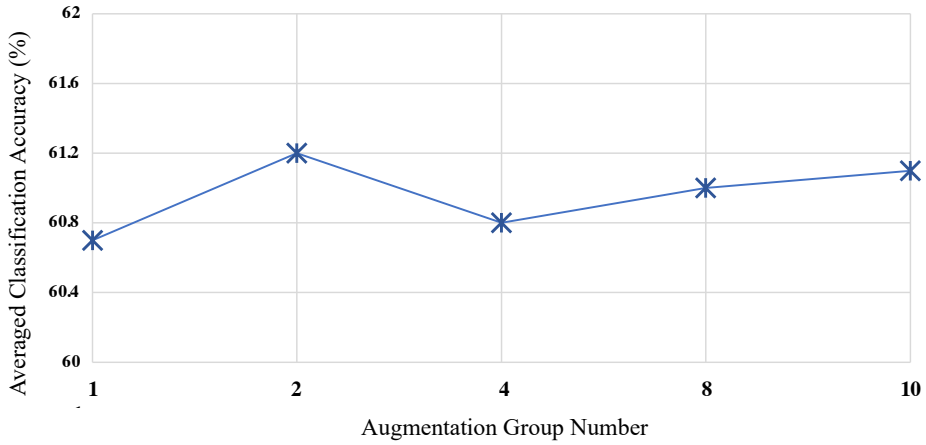


Figure 4.8: Averaged classification accuracy of SoTa-DiT with different augmentation group numbers M on ImageNet-C. The result shows that $M = 2$ yields the highest classification accuracy.

reason is twofold. When τ is small, the contrastive loss converges to 0 easily. For instance, if $\tau \to 0$, the positive pair is only slightly more similar than other samples, and the contrastive loss is close to 0 without proper training. Consequently, the effectiveness of the source and target contrastive loss diminishes. When τ surpasses 0.5, a larger τ makes it challenging to converge. As a result, TP and SP are prone to overfitting. Specifically, when τ is set to 1, if the positive pair similarity is 1 and the negative pair similarity is -1, the contrastive loss still deviates significantly from 0.

Second, the effect of τ_t is stronger than τ_s. We infer the reason is that the SP embedding is more stable with the source similarity loss, even when it is not properly supervised with the contrastive loss.

4.3.4.8 Augmentation Group Number M

We evaluate the influence of the augmentation groups M for the target contrastive learning. The results are depicted in Figure 4.8. We draw the following observations.

First, when M is 1, we observe a lower classification accuracy. We infer that when M equals 1, the data is insufficient for contrastive learning. Consequently, TP cannot absorb the target knowledge effectively.

Second, the accuracy remains relatively stable as M exceeds 2. We infer that 2 groups of augmented data are enough for contrastive learning. More augmented groups do not enhance the efficiency of target knowledge extraction; instead, they introduce

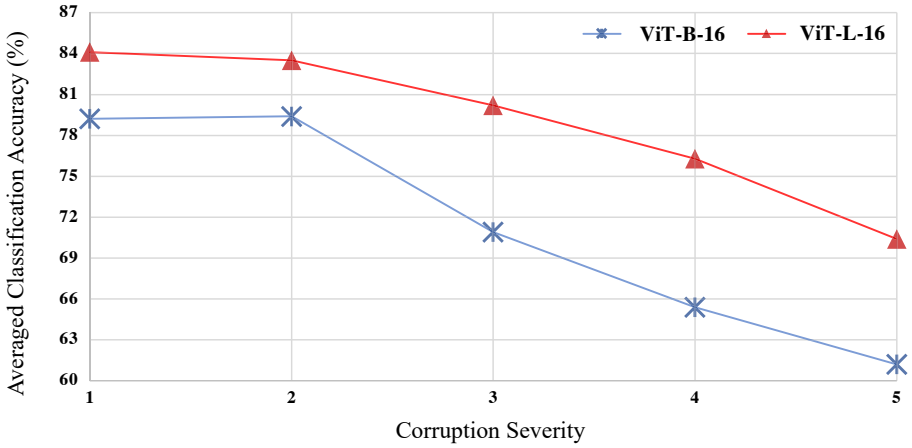


Figure 4.9: Averaged classification accuracy of SoTa-DiT with different corruption severity levels on ImageNet-C. The result shows that the higher the corruption level, the lower the classification accuracy. The reason is that higher-level corruption brings a large domain gap and is hard to learn at test time.

unnecessary computational load. As a result, we set M as 2 by default.

4.3.5 Observations

4.3.5.1 Corruption Severity

We test SoTa-DiT on ImageNet-C using five different corruption severities. Notice that all other experiments on ImageNet-C are conducted on severity level 5, the strongest corruption level. The average accuracies of SoTa-DiT with two types of baseline, namely ViT-B-16 and ViT-L-16, are depicted in Figure 4.9 with blue and red lines, respectively.

The result shows that, as the corruption severity increases, the average classification accuracy of SoTa-DiT with two backbones decreases. The reason is that as the severity of corruption increases, the test images become more dissimilar from the original images. As a result, the domain gap gets wider, making classification more challenging for the model.

4.3.5.2 Source Knowledge Preservation

In this and the following sections, we conduct experiments to substantiate our knowledge-disentangling claims further. First, we demonstrate that the source knowledge is well-preserved within the source prompt. Next, we demonstrate that the target prompt

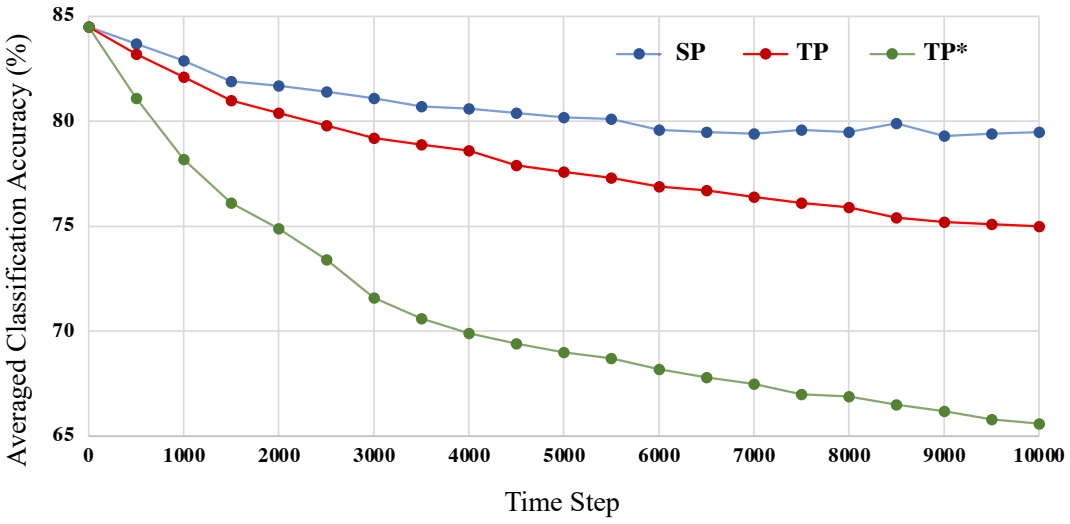


Figure 4.10: We evaluate the model directly on the original ImageNet test set at different time steps using 1) SP output only, 2) TP output only, and 3) TP output from a model without source prompt, denoted as TP∗. The result shows that the SP achieves the highest source accuracy, indicating that the source knowledge is preserved within SP.

efficiently extracts the target knowledge. All the experiments are conducted on the ImageNet-C dataset with the ViT-B-16 backbone.

We evaluate the model directly on the original ImageNet test set at different time steps to demonstrate that the source knowledge is preserved within the source prompt and adapted to other parts of the model. The classification accuracies at different time steps are depicted in Figure 4.10. Specifically, the blue, red, and green lines represent the prediction accuracy of 1) SP output, 2) TP output, and 3) TP output from a model without SP (denoted as TP∗), respectively. Based on the result, we draw the following observations:

First, SP successfully preserves the source knowledge. Comparing the classification accuracy of SP output and TP output, we see that the classification accuracy of SP outputs decreases much more slowly than that of TP output before the time step 6000. After 6000, the classification accuracy of SP output remains stable. This phenomenon proves that SP successfully extracts and preserves the source knowledge from the source model.

Second, SP adapts the source knowledge and helps the model retain it. Comparing the classification accuracy of TP output and TP∗ output, we see that the classification accuracy of TP output decreases much more slowly than that of TP∗ output. This

Table 4.5: Average domain classification accuracy of SP output and TP output. The result shows that the TP output yields significantly higher domain classification accuracy, indicating the TP output contains more information related to the target domain. This result proves that the TP mainly extracts the target-related knowledge.

Method	Average Acc.
SP	12.9
TP	27.4

phenomenon indicates that with SP, the TP also retains the ability to classify the source data better, proving that SP helps the model preserve the source knowledge.

Overall, the results prove that SP successfully extracts the source knowledge and prevents the model from forgetting.

4.3.5.3 Target Knowledge Extraction

We first train the SoTa-DiT model with only the target prompt and an additional domain classification head to distinguish different test domains. Then, we incorporate a domain entropy loss to train the classification head. Note that domain entropy loss only tunes the domain classification head. Each test image, along with its domain label, is fed into the network. After training the network with all 15 corruptions on ImageNet-C, we take the domain classification head for further evaluation.

We then train a standard SoTa-DiT model. At each time step, the domain classification accuracy of the model is evaluated with the aforementioned domain classification head. The average classification accuracy is shown in Table 4.5, where we compare the domain classification accuracy of 1) SP output and 2) TP output.

The result shows that the TP output achieves 27.4% accuracy while the SP output achieves only 12.9%. This result indicates that the TP embedding is more domain-distinguishable, indicating that TP extracts more target-relevant knowledge. Thus, it proves our knowledge disentangles claims and that TP extracts more target domain knowledge.

4.3.6 Limitation Discussion

The main limitation is that SoTa-DiT is only applicable to models with transformer layers. To use SoTa-DiT with the CNN backbones, one or a few transformer layers would need to be added.

Another limitation is that SoTa-DiT cannot be directly applied to small-scale datasets like CIFAR-10-C and CIFAR-100-C datasets. The reason for this is that directly training the ViT model on small-scale datasets without any pretraining does not yield satisfying results. The source model should be pre-trained on a large-scale dataset like ImageNet and then fine-tuned on small-scale datasets to achieve satisfying classification accuracy. As a result, we did not compare SoTa-DiT with other works on small-scale corruption datasets like CIFAR-10-C and CIFAR-100-C.

CHAPTER 5

SINGLE-SAMPLE CONTINUAL TEST-TIME ADAPTATION VIA EBAR: EFFICIENT BUFFER AND RESETTING

5.1 Introduction

Standard test-time adaptation (TTA) mitigates the domain gap between a source and a target domain [19, 65, 77, 79, 83, 100, 128, 170, 182, 209, 211]. Source-free TTA adapts a model with unlabelled target data only [100], distant from supervised [52, 98, 116] or semi-supervised [8, 101, 208, 210] adaptation. A challenging scenario for source-free TTA is single-sample TTA (STTA) [6, 38, 49, 117, 120, 149, 182, 211], where only one target sample is available at a time. Since target data can be highly noisy, the model can easily misestimate the distribution of the target domain. Such misestimation leads to severe parameter perturbation that increases the instability of adaptation[128, 149]. Meanwhile, a model may also encounter constant change in target domains. For instance, different weather and road conditions constantly affect an obstacle detection model in autonomous driving. As a result, continual test-time adaptation (CoTTA) [105, 106, 117, 151, 175] is introduced. In CoTTA, a model is adapted to dynamic environments over a long time, making it difficult for the model to remember the source knowledge and causing catastrophic forgetting. When applying a model in the real world, STTA and CoTTA can happen jointly. In such circumstances, a model is adapted to constantly changing target domains with a single sample per time for a long

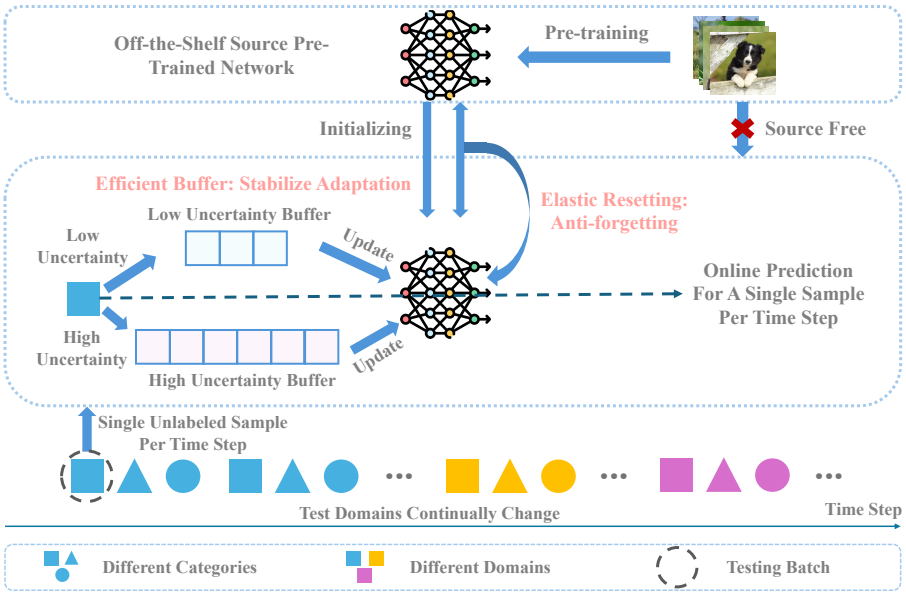


Figure 5.1: The concept of single-sample continual test-time adaptation (S-CoTTA) and our solution **E**fficient **B**uffer **a**nd **R**esetting (EBaR). In S-CoTTA, a model is adapted to target domains that constantly change over time, with only one sample per time step. Under such settings, a model usually suffers from 1) unstable adaptation and 2) catastrophic forgetting. To address this, we propose EBaR, which includes an efficient buffer and an elastic resetting unit. The efficient buffer consists of a low and a high uncertainty buffer that store low-uncertainty and high-uncertainty samples, respectively. The samples stored in two buffers update the model with different losses to stabilise tuning. The elastic resetting resets each parameter differently based on its sensitivity to domain shift, preventing forgetting while retaining necessary target domain knowledge.

time [182], posing two challenges: 1) efficient and stable adaptation with less memory overhead and timely prediction and 2) effective prevention of catastrophic forgetting.

To systematically address the aforementioned challenges, we introduce a **S**ingle-sample **Co**ntinual **T**est-**T**ime **A**daptation (S-CoTTA) task, as illustrated in Figure 5.1. In S-CoTTA, a pre-trained model is adapted to target domains that change continually over time using one unlabelled target data sample at a time. The primary objectives for S-CoTTA are twofold: 1) **Adaptation Stabilization**: Efficiently stabilising the adaptation of the model using a single target sample at a time, and 2) **Forgetting Prevention**: Effectively remembering the source knowledge over a long time. Recent works primarily address one of them. For instance, some apply a moving window or memory bank to a CoTTA method [117, 151, 175]. Although forgetting is mitigated, these methods failed to stabilise adaptation efficiently, and a large memory overhead is required. Others stabilise

adaptation by filtering out high-uncertainty samples [149, 202]. Although stabilising the adaptation, these methods fail to prevent forgetting. We contend that achieving both efficient adaptation stabilising and effective prevention of forgetting is essential.

To this end, we propose a novel **E**fficient **B**uffer **a**nd **R**esetting (EBaR) method to address the challenges posed by S-CoTTA. As illustrated in Figure 5.1 and Figure 5.2, EBaR consists of two key components: an efficient buffer and an elastic resetting control unit.

First, EBaR incorporates a novel memory-efficient buffer, named efficient buffer (EB), to stabilise the adaptation progress with less extra memory overhead. EB consists of two buffers: a low uncertainty buffer (LUB) and a high uncertainty buffer (HUB). During adaptation, our model first predicts the category of a sample directly. Then, a symmetric entropy weight is calculated to indicate the uncertainty of the sample [117, 128], and the sample is categorised into high- and low-uncertainty levels with a predefined threshold. Samples with low uncertainty are stored within the LUB of a smaller capacity; others are stored within the HUB of a larger capacity. When full, buffers release samples to update the model. A soft likelihood ratio and a symmetric entropy loss [38] are calculated for LUB samples, while a weighted contrastive learning loss is computed for HUB samples. EB brings two benefits. First, it congregates samples and applies different losses according to sample uncertainty to stabilise adaptation. Second, compared with moving windows and other memory banks, EB is smaller in size. Thus, it requires less memory overhead and allows timely updates, achieving higher efficiency.

Second, inspired by [139], EBaR incorporates a novel elastic resetting (ER) control unit to prevent catastrophic forgetting. ER adopts a reset clock and a dynamic weight reset for each parameter. At each time step, the reset clock increases by 1) a fixed amount plus 2) a value proportional to the updating magnitude of the parameter. When the reset clock exceeds a threshold, the dynamic weight reset is applied to the corresponding parameter based on its sensitivity to domain shift. Then, the reset clock is set back to zero. Elastic resetting ensures that parameters that are more susceptible to domain shift are reset more frequently and closer to the initial value. ER has two benefits. First, it effectively mitigates catastrophic forgetting. Second, compared to [139], where a whole model is reset after a fixed number of time steps, our strategy allows the model to retain target knowledge by resetting insensitive parameters less frequently.

Finally, comprehensive experiments are conducted to evaluate the effectiveness of EBaR. We demonstrate that in EBaR: 1) the efficient buffer significantly improves classification accuracy with reduced extra memory overhead compared to other counterparts

like the moving window, and 2) the elastic resetting enables the model to achieve higher classification accuracy stably over a long period of time. Combining the benefits of the two key components, EBaR achieves SOTA performance across multiple corruption datasets, including CIFAR10-C, CIFAR100-C, ImageNet-C, and CCC under the S-CoTTA setting with less extra memory overhead.

The contributions of EBaR can be summarized as follows:

• We introduce a challenging **S**ingle-sample **Co**ntinual **T**est-**T**ime **A**daptation (S-CoTTA) task. In S-CoTTA, a model is adapted to target domains that change constantly over time using a single unlabelled sample per time step. The objective is to achieve stable, efficient adaptation with one sample at a time and prevent catastrophic forgetting over a long time.

• We proposed a novel **E**fficient **B**uffer **a**nd **R**esetting (EBaR) method for S-CoTTA. To achieve stable adaptation, we introduce a novel memory-efficient buffer named the efficient buffer, consisting of a smaller and a larger buffer to store samples with low and high uncertainty levels, respectively. The stored samples are then used to update the model with different losses. To prevent catastrophic forgetting, we introduce a novel elastic resetting control unit to reset each parameter differently based on its sensitivity to domain shifts.

• We conduct comprehensive experiments to evaluate the effectiveness of the efficient buffer and the elastic resetting in EBaR. The results demonstrate the effectiveness of both components.

5.2 Methodology

5.2.1 Problem Formulation

This paper focuses on the image classification task under the source-free single-sample continual test-time adaptation (S-CoTTA) setting. We begin with a source model $f_{\Theta_S}(x)$ pre-trained on the initial source domain Φ_S with data and labels $(\mathscr{X}^S, \mathscr{Y}^S)$, where $(\mathscr{X}^S, \mathscr{Y}^S)$ is unavailable for adaptation when testing. The goal is to achieve a higher classification accuracy for images from target test domains. During the test time, the model, initialized with $f_{\Theta_S}(x)$, classifies the target batches $\mathscr{X}^T = (X_1, X_2, ...)$ from test domains Φ_T different from Φ_S. In S-CoTTA, each data batch X_t contains only one sample, and the distribution of X_t changes continually. Thus, we have $X_t = (x_t)$. The data batches are presented to the model following a time-step sequence. At each time step t, the

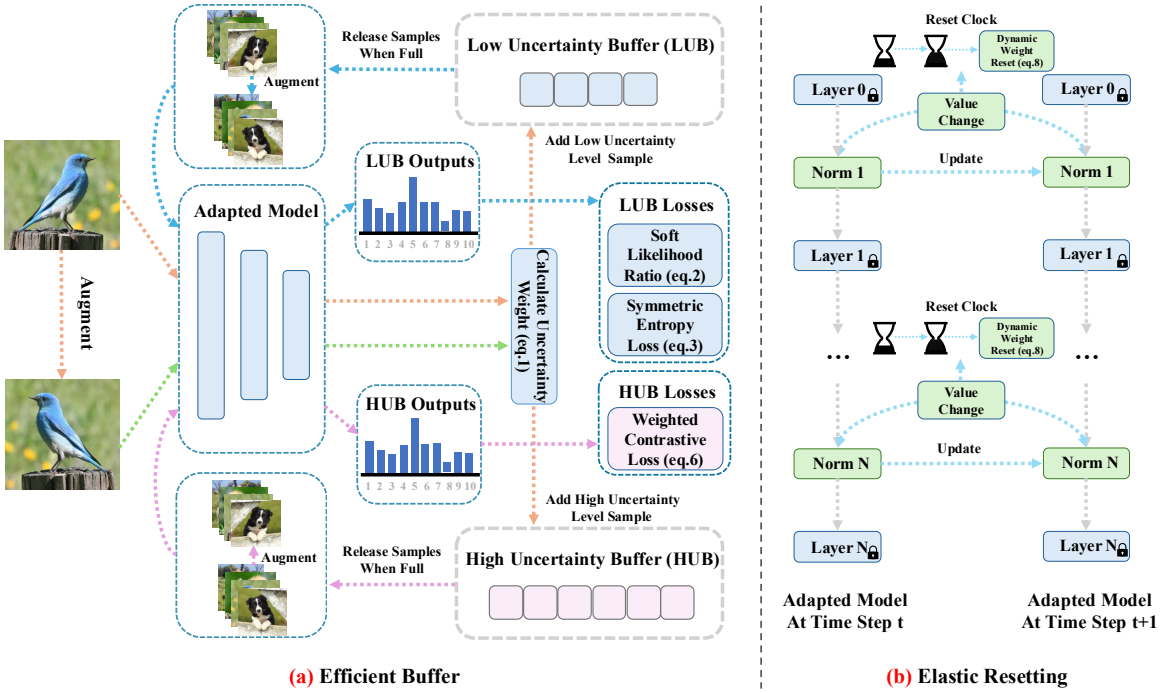


Figure 5.2: EBaR, designed for the S-CoTTA task, consists of two key components: **(a)** an Efficient Buffer (EB) and **(b)** an Elastic Resetting (ER). As shown in **(a)**, EB comprises a smaller Low Uncertainty Buffer (LUB) and a larger High Uncertainty Buffer (HUB). The original and augmented samples are first classified. Based on that, the uncertainty level for each sample is evaluated. The low-uncertainty and high-uncertainty samples are stored in LUB and HUB, respectively. Once a buffer is full, samples stored in the buffer are augmented and fed into the model for outputs. LUB and HUB losses are then computed for samples stored in the corresponding buffer and backpropagated to update the model. As shown in **(b)**, ER features a reset clock that tracks parameter changes, enabling an asynchronous reset for each parameter based on its sensitivity to domain changes. When the clock exceeds the threshold, a dynamic weight reset is applied to set a parameter closer to its initial value when it is sensitive.

adapted model $f_{\Theta_t}(\mathrm{x})$ classifies the sample x_t while being adapted to Φ_T using unlabelled x_t for the next time step $t+1$.

5.2.2 Overall Architecture

The overall architecture of our solution, named **E**fficient **B**uffer **a**nd **R**esetting (EBaR), is illustrated in Figure 5.2. We begin with an initial source model, $f_{\Theta_S} = (F_S, H_S)$, and an adapted model, $f_{\Theta_t} = (F_t, H_t)$, at a time step t. Here, F and H represent an image encoder and a classification head with a softmax layer, respectively. At the time step 0,

f_{Θ_0} is initialized with f_{Θ_S}.

At time step t, a test data batch $X_t = (x_t)$,consisting of a single sample x_t is presented to the adapted model $f_{\Theta_t} = (F_t, H_t)$ for classification. x_t belongs to one of the C categories, and the goal of f_{Θ_t} is to predict the category. First, x_t is fed into the image encoder F_t. We denote the encoded image as $o_t = F_t(x_t)$. Then, o_t is fed into the classification head H_t to generate the prediction $y_t = H_t(o_t)$. Finally, the classification accuracy is computed for x_t using y_t.

Meanwhile, the adapted model f_{Θ_t} is updated for the next time step $t+1$ using o_t and y_t. In EBaR, only the normalisation layers are opened for updating, with other parameters fixed. The update procedure incorporates the **E**fficient **B**uffer (EB) and **E**lastic **R**esetting (ER) control unit, both of which are introduced with details in the following subsections.

5.2.3 Efficient Buffer for Stable Adaptation

Algorithm 1: Update the model with EB

Input: x_t, LUB, HUB
$x'_t \leftarrow$augmentation(x_t);
$y_t \leftarrow f_{\Theta_t}(x_t)$, $y'_t \leftarrow f_{\Theta_t}(x'_t)$;
$\mathscr{L}_{Ut} \leftarrow \mathscr{L}_U(y_t, y'_t)$;
if $\mathscr{L}_{Ut} < \epsilon$ **then**
  if *number of samples in LUB equals to* M **then**
    $X_L \leftarrow$ LUB.stored_samples();
    $\mathscr{L}_{Lt} \leftarrow \mathscr{L}_L(X_L)$;
    LUB.clear_stored_samples();
  end
  LUB.add_sample(x_t);
else
  if *number of samples in HUB equals to* N **then**
    $X_H \leftarrow$ HUB.stored_samples();
    $\mathscr{L}_{Ht} \leftarrow \mathscr{L}_H(X_H)$;
    HUB.clear_stored_samples();
  end
  HUB.add_sample(x_t);
end
$\mathscr{L}_t \leftarrow \mathscr{L}_{Lt} + \mathscr{L}_{Ht}$;
$\mathscr{L}_t$.backpropagation();

EBaR introduces a novel memory-efficient multilevel buffer named **E**fficient **B**uffer

(EB). The purpose of EB is twofold: 1) stabilising the adaptation by categorising the samples based on their uncertainty level and applying different losses, and 2) reducing the memory overhead and enabling timely updates.

Specifically, EB comprises a smaller **L**ow **U**ncertainty **B**uffer (LUB) and a larger **H**igh **U**ncertainty **B**uffer (HUB). LUB has a capacity of M samples, while HUB can store up to N samples. Samples with different uncertainty levels are stored in different buffers. Following [128, 149], we apply a symmetric entropy loss to indicate uncertainty levels for samples. Given a sample x_t, we obtain the prediction $y_t = f_{\Theta_t}(x_t)$. Then, the sample is augmented by applying colour jitter, affine transformations, and horizontal flipping. The augmented sample, denoted as x'_t, is then fed into the model to obtain the prediction $y'_t = f_{\Theta_t}(x'_t)$. The uncertainty weight, denoted as $\mathscr{L}_U$, is computed as follows:

$$\mathscr{L}_U(y_t, y'_t) = -\big(\sum_{c=1}^{C} y_{tc} \log y'_{tc} + \sum_{c=1}^{C} y'_{tc} \log y_{tc}\big). \tag{5.1}$$

Next, EBaR updates the backbone model using LUB and HUB, as outlined in Algo.1, where ϵ is a predefined threshold.

5.2.3.1 Low Uncertainty Buffer

The **L**ow **U**ncertainty **B**uffer (LUB) stores samples with low uncertainty levels. LUB has a smaller capacity M because low-uncertainty samples contain high-quality target knowledge with less noise [128] and can stably update the model without perturbation, even within a considerably small batch.

In Algo.1, LUB stores the samples whose $\mathscr{L}_U$ are below the threshold ϵ. Once the number of samples stored in LUB reaches M, the stored samples $X_L = (x_1, x_2, ...x_M)$ are released and processed to calculate the LUB loss $\mathscr{L}_L$. $\mathscr{L}_L$ consists of a soft likelihood ratio loss (SLR) $\mathscr{L}_{SLR}$, and a symmetric entropy loss (SEL) $\mathscr{L}_{SEL}$.

Specifically, EBaR first obtains the predictions of X_L, denoted as $Y_L = (y_1, y_2, ..., y_M)$. $\mathscr{L}_{SLR}$ is calculated as follows:

$$\mathscr{L}_{SLR}(Y_L) = -\frac{1}{M} \sum_{m=1}^{M} \sum_{c=1}^{C} y_{mc} \log\left(\frac{y_{mc}}{\sum_{j \neq c}^{C} y_{mj} + \xi}\right). \tag{5.2}$$

Meanwhile, EBaR augments X_L to generate augmented samples $X'_L = (x'_1, x'_2, ...x'_M)$. X'_L is then fed into f_{Θ_t} to obtain the prediction $Y'_L = (y'_1, y'_2, ..., y'_M)$. $\mathscr{L}_{SEL}$ is then calculated as follows:

$$\mathscr{L}_{SEL}(Y_L, Y'_L) = -\frac{1}{2M}\sum_{m=1}^{M}(\sum_{c=1}^{C} y_{mc}\log y'_{mc} + \sum_{c=1}^{C} y'_{mc}\log y_{mc}). \tag{5.3}$$

$\mathscr{L}_{SLR}$ is less dominated by the predictions with low confidence [117, 122] and $\mathscr{L}_{SEL}$ is more robust to label noise compared to the entropy loss [38]. They both reduce the potential impact of pseudo-label perturbation in self-supervised adaptation. Jointly using two losses further stabilises model updates when very limited samples are available at a time. Finally, the LUB loss $\mathscr{L}_L$ is calculated as:

$$\mathscr{L}_L = \mathscr{L}_{SLR} + \beta\mathscr{L}_{SEL}. \tag{5.4}$$

5.2.3.2 High Uncertainty Buffer

The **H**igh **U**ncertainty **B**uffer (HUB) stores the samples with high uncertainty levels. These samples disturb the adaptation, especially in a small batch [128]. However, we argue that completely discarding these samples, a strategy adopted by [128, 149], is suboptimal. Our rationale is twofold: 1) the model may initially overestimate uncertainty levels and can refine them after a few updates, and 2) high-uncertainty samples also contain valuable target-domain knowledge. To effectively exploit these samples, we allocate a larger capacity N to HUB. This setting reduces the immediate impact of these high-uncertainty samples and allows the model to reassess their uncertainty after a few more updates.

In Algo.1, HUB stores the samples whose $\mathscr{L}_U$ are higher than the threshold ϵ. Once the number of samples stored in HUB reaches N, the stored samples $X_H = (x_1, x_2, ..., x_N)$ are released and processed to calculate the HUB loss $\mathscr{L}_H$. The loss $\mathscr{L}_H$ is calculated based on an entropy weight $\Delta_H = (\delta_1, \delta_2, ...\delta_n, ...\delta_N)$ and a weighted contrastive learning loss (CLL) $\mathscr{L}_{CLL}$.

Specifically, EBaR first feeds X_H into f_{Θ_t} to obtain the predictions $Y_H = (y_1, y_2, ...y_N)$. Next, augmentation is applied to X_H to generate augmented samples $X'_H = (x'_1, x'_2, ...x'_N)$. Then, X'_H is fed into f_{Θ_t} to generate predictions $Y'_H = (y'_1, y'_2, ...y'_N)$. The entropy weight δ_n for each stored sample x_n is computed as:

$$\delta_n = \exp\big(min(\sum_{c=1}^{C} y_{nc}\log y'_{nc} + \sum_{c=1}^{C} y'_{nc}\log y_{nc} + \epsilon, 0)\big). \tag{5.5}$$

Meanwhile, X_H and X'_H are fed into the encoder F_t to obtain the encoded images $O_H = (o_1, o_2, ...o_N)$ and $O'_H = (o'_1, o'_2, ...o'_N)$, respectively. $\mathscr{L}_{CLL}$ is then computed as:

$$\mathrm{L}_{CLL}(O_H, O'_H, \Delta_H) = -\frac{1}{N}\sum_{n=1}^{N}\delta_n\left(\sum_{c=1}^{C}\frac{\exp\left(\mathrm{sim}(o_{nc}, o'_{nc})/\tau\right)}{\sum_{k=1}^{N}\exp\left(\mathrm{sim}(o_{nc}, o'_{kc})/\tau\right)}\right), \tag{5.6}$$

where sim represents the cosine similarity, and τ is the temperature.

By weighting contrastive learning loss with the entropy weight, EBaR achieves two key objectives: 1) extracting target-domain knowledge from samples with higher uncertainty without using potentially noisy pseudo-labels and 2) re-estimating sample uncertainty levels to modulate their contribution to model updates. This strategy mitigates the adverse effects of samples with high uncertainty while extracting useful target-domain knowledge. The HUB loss $\mathscr{L}_H$ is computed as:

$$\mathscr{L}_H = \mathscr{L}_{CLL}. \tag{5.7}$$

5.2.4 Elastic Resetting for Anti-forgetting

Algorithm 2: Elastic resetting for parameters

Input: Θ_t, Θ_{t-1}, R, μ, η, Υ, T
for $(\theta_t,\ \theta_{t-1})$ *in* $(\Theta_t,\ \Theta_{t-1})$ **do**
  if θ_t *is updated* **then**
    $D \leftarrow \frac{||\theta_t - \theta_{t+1}||}{||\theta_t|| + ||\theta_{t+1}||}$;
    $R_\theta \leftarrow R_\theta + \eta D + \mu$;
  else
    $R_\theta \leftarrow R_\theta + \mu$;
  end
  $T_\theta \leftarrow T_\theta + \mu$;
  if $R_\theta \geq \Upsilon$ **then**
    θ.reset(T_θ, Υ) ;
    $R_\theta \leftarrow 0$, $T_\theta \leftarrow 0$;
  end
end

EBaR incorporates a **E**lastic **R**esetting (ER) control unit to mitigate catastrophic forgetting over long time steps. Unlike existing strategies [139, 152], which periodically reset the entire model back to the initial state after a fixed number of time steps, ER employs an asynchronous resetting mechanism. In ER, different parameters are reset at varying frequencies based on their sensitivity to domain shifts. Our approach is

motivated by two key insights: 1) the resetting operation effectively restores the source knowledge, and 2) not all parameters are equally susceptible to domain shifts: some are comparably robust to perturbations brought by domain shifts and effectively absorb rich target-domain knowledge. ER balances preserving source knowledge and retaining target knowledge by more frequently resetting only the parameters sensitive to perturbations.

To effectively implement ER, EBaR uses a novel reset clock R for each parameter θ, which operates based on Algo.2. In the algorithm, Θ_t and Θ_{t-1} represent the values of the updated parameters at the time steps t and $t-1$, respectively. The reset process begins at time step 1, with three hyper-parameters: μ, η, and Υ. In Algo.2, as the time step increases, the reset clock R_θ for each parameter θ increases by: 1) a fixed increment μ, and 2) an increment related proportionally to the magnitude of the changes in the parameter value during the update, adjusted by η. When R_θ reaches or exceeds the threshold Υ, the corresponding parameter θ is reset.

Unlike [139], which resets parameters to their initial values θ_0, ER employs a novel dynamic weight reset based on the sensitivity of each parameter to perturbations. EBaR also incorporates a sensitive indicator T in Algo.2. Every time step, T_θ for each parameter θ increases by μ. The dynamic weight reset is performed as follows:

$$\theta'_t = (1-\alpha)\frac{T_\theta}{\Upsilon}\theta_t + (1-\frac{T_\theta}{\Upsilon}+\alpha\frac{T_\theta}{\Upsilon})\theta_0, \tag{5.8}$$

where α is a hyper-parameter that adjusts the reset.

When a reset occurs, a higher T_θ indicates that a parameter is reset less frequently and is less sensitive to perturbations caused by domain shifts. Consequently, parameters with higher T_θ are reset closer to their initial values θ_0, ensuring a balance between retraining target knowledge and mitigating forgetting.

5.3 Experiments

5.3.1 Settings

5.3.1.1 Datasets

EBaR is evaluated on four CoTTA datasets: CIFAR10-C, CIFAR100-C, ImageNet-C (ImageNet-C) [68, 69, 136, 147], and CCC [139]. CIFAR10-C, CIFAR100-C, and ImageNet-C contain 15 types of image corruption, each with 5 severity levels. The experiments are conducted at severity level 5. The CCC dataset (Continuously Changing Corruptions)

Table 5.1: The averaged classification accuracy (%) of different methods. All the methods except EBaR have a moving window of size 32 added for better convergence. 'Source' represents the backbone model without adaptation. 'CMAE' only supports Transformer baselines. EBaR achieves SOTA performance across multiple datasets with different baselines.

Dataset	Backbone	SOURCE	TENT [170]	COTTA [175]	ETA [128]	ViDA [106]	REALM [149]	SAR [129]	IST [115]	CMAE [105]	ROTTA [202]	RDUMB [139]	BDG [197]	ROID [117]	EBaR
CIFAR10	WideResNet	56.4	73.8	76.6	77.4	79.9	79.8	77.1	**80.8**	-	79.0	79.9	80.4	80.6	**84.3** (+3.5)
CIFAR100	ResNeXT-29	53.6	29.8	61.6	59.7	66.5	65.3	67.1	66.2	-	64.8	66.0	**68.0**	**67.6**	**71.9** (+3.9)
ImageNet-C	ResNet-50	18.1	33.1	33.7	39.6	43.8	41.7	36.6	36.2	-	30.1	40.3	**44.1**	43.7	**47.6** (+3.5)
	ViT-B-16	39.9	23.2	13.7	43.6	50.5	44.9	49.8	50.4	51.3	44.1	46.1	52.4	**52.5**	**56.7** (+4.2)
	Swin-B	45.3	19.8	9.9	54.8	56.8	52.7	47.8	60.2	**62.1**	44.8	53.2	**62.1**	61.8	**67.3** (+5.2)
CCC	ResNet-50	33.5	5.7	7.6	6.8	16.8	31.5	39.3	15.8	-	10.7	41.3	45.2	**47.6**	**50.6** (+3.0)
	ViT-B-16	53.5	4.8	5.3	9.0	36.9	62.8	**65.8**	39.7	30.9	9.7	64.9	64.2	65.6	**70.3** (+4.5)
	Swin-B	55.8	3.9	4.1	7.9	39.8	63.9	65.4	39.6	30.2	9.8	64.5	64.1	**65.5**	**70.9** (+5.4)

comprises 7.5 million images with smooth transitions between different ImageNet-C corruptions [139]. CCC is designed to assess the stability of test-time adaptation methods over extended time steps. Our experiments use CCC-40, where the source ResNet-50 model achieves 35% to 40% accuracy without adaptation. The results on the CCC datasets are averaged over 3 different random seeds.

5.3.1.2 Implementation Details

EBaR is evaluated on five backbones: WideResNet-28 (WRN-28) [204], ResNeXt-29 [189], ResNet-50 [64], ViT Base (Vit-B-16) [39], and Swin Transformer Base (Swin-B-16) [112]. Following [117], the evaluation on CIFAR10-C is performed with WRN-28, and CIFAR100-C with ResNeXt-29. ResNet-50, Vit-B-16, and Swin-B-16 are evaluated for ImageNet-C and CCC. WRN-28 and ResNeXt-29 are pretrained with CIFAR datasets [87], while ResNet-50, Vit-B-16 and Swin-B-16 are pretrained with ImageNet [34], following [24, 155, 164, 207, 221]. We set the learning rate to 1×10^{-4}, LUB size M to 2, HUB size N to 8, ϵ (Algo.1) to 6.2, ξ (Eq.5.2) to 0.01, β (Eq.5.4) to 1, and τ (Eq.5.6) to 0.05, μ (Algo.2) to 1, η (Algo.2) to 100, Υ (Algo.2) to 8000, and α (Eq.5.8) to 0.99 as default.

5.3.2 Effectiveness of EBaR

We compare EBaR with SOTA CoTTA and STTA methods in Table 5.1 and Table 5.2. For a fair comparison, we apply a moving window of size 32 to all counterpart methods, following [38, 117, 128].

Table 5.2: The classification accuracy (%) of different methods using ResNet-50 on ImageNet-C across 15 corruptions. EBaR achieves the highest accuracy on 10 out of 15 corruptions.

Method	gaussian	shot	impulse	defocus	glass	motion	zoom	snow	frost	fog	bright	contrast	elsatic	pixelate	jpeg	**Avg.**
SOURCE	2.2	3.0	1.9	18.6	10.2	14.8	22.0	16.5	23.2	24.1	58.7	6.0	17.5	20.7	31.5	18.1
TENT	15.1	21.4	23.2	17.8	23.1	29.9	40.8	33.5	32.0	43.4	53.4	24.0	45.8	49.1	43.3	33.1
COTTA	12.8	17.1	19.7	17.0	17.6	24.2	37.5	34.1	32.6	47.7	59.9	30.6	48.6	55.4	51.2	33.7
ETA	21.4	31.9	34.6	21.7	26.9	35.1	44.0	39.8	37.4	50.3	59.6	36.4	50.6	53.6	50.7	39.6
ViDA	18.8	23.3	24.7	22.0	21.7	33.6	42.4	38.7	36.1	48.2	59.5	32.0	47.6	53.1	50.7	36.8
REALM	22.4	30.1	35.0	24.1	25.2	36.2	44.3	44.3	43.2	45.2	**73.1**	**52.8**	40.2	55.1	54.2	41.7
SAR	17.9	23.8	27.7	20.1	23.6	33.9	43.1	37.0	37.1	47.0	60.2	29.8	49.1	50.2	48.1	36.6
IST	27.3	28.6	26.9	20.0	18.8	29.8	41.9	39.7	35.9	49.7	60.8	15.8	47.2	51.3	49.0	36.2
ROTTA	10.7	15.5	14.7	7.8	15.3	23.8	38.2	32.1	33.6	43.1	60.5	21.2	41.7	49.2	44.4	30.1
RDUMB	21.8	33.2	35.8	22.4	25.8	37.0	44.8	41.4	37.1	49.8	59.6	38.1	51.5	54.2	51.4	40.3
BDG	24.8	33.4	35.7	28.9	**39.5**	38.9	50.6	46.3	**45.1**	56.4	62.5	39.6	54.9	54.5	50.8	44.1
ROID	26.3	**37.8**	36.9	29.7	32.1	42.9	46.6	43.8	43.7	53.2	63.8	37.4	52.7	56.3	52.0	43.7
EBaR (ours)	**30.9**	37.5	**41.5**	**32.8**	33.2	**43.9**	**53.4**	**49.8**	43.9	**59.6**	69.9	43.6	**56.9**	**59.8**	**56.8**	**47.6**

The results show that under S-CoTTA, EBaR outperforms the SOTA methods by clear margins: (1) across multiple datasets with various backbones, as shown in Table 5.1, and (2) on 10 of 15 corruptions for the ImageNet-C dataset using the ResNet-50 backbone, as shown in Table 5.2. For instance, EBaR achieves +3.5%, +4.2%, and +5.2% on ImageNet-C with ResNet-50, ViT-B-16, and Swin-B-16. These results demonstrate the superiority of EBaR on the S-CoTTA task.

5.3.3 Ablation Studies of Key Components

5.3.3.1 Efficiency Buffer

We evaluate the effectiveness of the Efficient Buffer, including LUB, HUB, and their associated loss functions. For comparison, we choose the baseline method ETA [128] with a moving window size of 32.

As shown in Table 5.3, both LUB and HUB, along with their loss functions, are beneficial to S-CoTTA and improve accuracy. For instance, using LUB alone brings a +2.5% accuracy gain, and using HUB alone brings a +1.8% accuracy gain. Both loss functions in LUB contribute to the increase. These results indicate that the efficient buffer stabilises the adaptation and increases the accuracy.

Furthermore, combining LUB and HUB yields a more significant accuracy gain (+5.6%) compared to using them separately (+2.5% and +1.8%). This result indicates that the information within the low- and high-uncertainty samples is complementary, and combining them is more beneficial than using only the low-uncertainty samples.

Table 5.3: Evaluation of the key components: efficient buffer and elastic resetting on ImageNet-C with ResNet-50. We list the averaged accuracy (%). 'DWR' represents dynamic weight reset (Eq.5.8). The result proves the effectiveness of both.

Method	$\mathscr{L}_{SLR}$	$\mathscr{L}_{SEL}$	$\mathscr{L}_{CLL}$	$Clock$	DWR	Average Acc.
Baseline (ETA)	✗	✗	✗	✗	✗	39.6
LUB + HUB	✓	✓	✓	✗	✗	45.2 (+5.6%)
LUB only	✓	✓	✗	✗	✗	42.1 (+2.5%)
HUB only	✗	✗	✓	✗	✗	41.4 (+1.8%)
LUB only -$\mathscr{L}_{SLR}$	✓	✗	✗	✗	✗	41.1 (+1.5%)
LUB only -$\mathscr{L}_{SEL}$	✗	✓	✗	✗	✗	40.2 (+0.6%)
ER only	✗	✗	✗	✓	✓	44.3 (+4.7%)
ER only - DWR	✗	✗	✗	✓	✗	42.7 (+3.1%)
DWR only	✗	✗	✗	✗	✓	42.2 (+2.6%)
EBaR	✓	✓	✓	✓	✓	**47.6 (+8.0%)**

5.3.3.2 Elastic Resetting

We evaluate the effectiveness of the Elastic Resetting and its two key components, the reset clock and the dynamic weight reset. As shown in Table 5.3, both the clock-based reset and the dynamic weight reset increase performance. For instance, using clock-based reset boosts the accuracy by 3.1%, and using the dynamic weight reset boosts the accuracy by 2.6%. These results show that both strategies are beneficial to S-CoTTa. Jointly applying both improves accuracy by 4.7%, proving that the two key components are complementary and bring cumulative benefits. Moreover, combining ER and EB yields a more significant increase in accuracy of +8.0%, proving that both ER and EB are critical in EBaR and bring cumulative benefits when used together.

5.3.4 Ablation Studies of Loss Functions

In this section, we justify our selection of the loss function in EBaR. The effectiveness of the loss functions is demonstrated by comparing them with other loss function options. Four key losses in EBaR are evaluated: uncertainty weight $\mathscr{L}_U$, soft likelihood ratio $\mathscr{L}_{SLR}$ for LUB, symmetric entropy loss $\mathscr{L}_{SEL}$ for LUB, and weighted contrastive learning loss $\mathscr{L}_{CLL}$ for HUB.

Table 5.4: Evaluation of the uncertainty weight $\mathscr{L}_U$ on ImageNet-C with ResNet-50. We list the averaged accuracy (%). Our strategy in EBaR is compared with standard entropy loss and diversity-based weight, proposed in [128]. We adjust the corresponding threshold ϵ for different selections and report the best result. The result shows that applying the symmetrical entropy loss as the uncertainty weight is superior.

Method	Average Acc.
Baseline (ETA)	39.6
Entropy Loss	45.9 (+6.3%)
Diversity-based Weight [128]	46.8 (+7.2%)
EBaR	**47.6 (+8.0%)**

Table 5.5: Evaluation of the soft likelihood ratio loss $\mathscr{L}_{SLR}$ for LUB on ImageNet-C with ResNet-50. We list the averaged accuracy (%). In this experiment, the counterparts replace the soft likelihood ratio with 1) a standard entropy loss and 2) a standard contrastive loss. The result shows that EBaR with the soft likelihood ratio loss achieves the highest accuracy, proving the effectiveness of $\mathscr{L}_{SLR}$.

Method	Average Acc.
Baseline (ETA)	39.6
Entropy Loss	46.3 (+6.7%)
Contrastive Loss	46.5 (+6.9%)
EBaR-$\mathscr{L}_{SLR}$	46.2 (+6.6%)
EBaR	**47.6 (+8.0%)**

5.3.4.1 Uncertainty Level Loss $\mathscr{L}_U$

We evaluate the effectiveness of the uncertainty weight $\mathscr{L}_U$. In EBaR, the symmetric entropy loss is selected as the uncertainty weight. In this experiment, different options are examined in EBaR. Here, we choose the standard entropy loss and the diversity-based weight [128] as counterparts. Furthermore, the uncertainty level threshold ϵ is adjusted to achieve the highest classification accuracy for different losses. The model with entropy loss achieves the highest accuracy when $\epsilon = 2.8$. Moreover, $\epsilon = 0.06$ is the optimum for diversity-based weight. As shown in Table 5.4, choosing symmetric entropy loss as uncertainty weight achieves the highest average accuracy, surpassing entropy loss and diversity-based weight by 1.7% and 0.8%, respectively.

5.3.4.2 Soft Likelihood Ratio $\mathscr{L}_{SLR}$

We evaluate the effectiveness of the soft likelihood ratio $\mathscr{L}_{SLR}$ for LUB. In this experiment, we adopt other loss functions to replace $\mathscr{L}_{SLR}$. Here, 1) in 'Entropy Loss', a

Table 5.6: Evaluation of the symmetric entropy loss $\mathscr{L}_{SEL}$ for LUB on ImageNet-C with ResNet-50. We list the averaged accuracy (%). This experiment replaces the symmetric entropy loss with 1) a standard entropy loss and 2) a standard contrastive loss. The result shows that EBaR with the symmetric entropy loss achieves the highest accuracy, proving the effectiveness of $\mathscr{L}_{SEL}$.

Method	Average Acc.
Baseline (ETA)	39.6
Entropy Loss	47.1 (+7.5%)
Contrastive Loss	47.2 (+7.6%)
EBaR-$\mathscr{L}_{SEL}$	47.1 (+7.5%)
EBaR	**47.6 (+8.0%)**

standard entropy loss replaces $\mathscr{L}_{SLR}$, 2) in 'Contrastive Loss', a standard contrasive learning loss, similar to $\mathscr{L}_{CLL}$ but without the sample weight Δ_H, replaces $\mathscr{L}_{SLR}$, and 3) in 'EBaR-$\mathscr{L}_{SLR}$', $\mathscr{L}_{SLR}$ is removed without replacement loss. As shown in Table 5.5, EBaR using soft likelihood ratio loss achieves the highest averaged accuracy, surpassing the counterparts with entropy loss, contrastive loss, and no soft likelihood ratio loss by 1.3%, 1.1%, and 1.4%, respectively.

5.3.4.3 Symmetric Entropy Loss $\mathscr{L}_{SEL}$

We evaluate the effectiveness of the symmetric entropy loss $\mathscr{L}_{SEL}$. Similarly to the previous experiment, other loss functions replace $\mathscr{L}_{SEL}$ in the counterparts. Here, 1) in 'Entropy Loss', a standard entropy loss replaces $\mathscr{L}_{SEL}$, 2) in 'Contrastive Loss', a standard contrasive learning loss, similar to $\mathscr{L}_{CLL}$ but without the sample weight Δ_S, replaces $\mathscr{L}_{SEL}$, and 3) in 'EBaR-$\mathscr{L}_{SEL}$', $\mathscr{L}_{SEL}$ is removed without replacement loss. As shown in Table 5.6, EBaR using symmetric entropy loss achieves the highest averaged accuracy, surpassing counterparts with entropy loss, contrastive loss, and no soft likelihood ratio loss by 0.5%, 0.4%, and 0.5%, respectively.

5.3.4.4 Contrastive Learning Loss $\mathscr{L}_{CLL}$

We evaluate the effectiveness of the weighted contrastive learning loss $\mathscr{L}_{CLL}$ for HUB. EBaR is compared with 1) 'Baseline (ETA)', the baseline ETA [128] model, 2) 'EBaR-$\mathscr{L}_{CLL}$', EBaR without $\mathscr{L}_{CLL}$ loss, 3) 'Entropy Loss', where a standard entropy loss replaces $\mathscr{L}_{CLL}$, 4) '$\mathscr{L}_{SLR}$', where $\mathscr{L}_{SLR}$ replaces $\mathscr{L}_{CLL}$, 5) '$\mathscr{L}_{SEL}$', where $\mathscr{L}_{SEL}$ replaces $\mathscr{L}_{CLL}$, 6) '$\mathscr{L}_{SLR}$+$\mathscr{L}_{SEL}$', where $\mathscr{L}_{SLR}$ and $\mathscr{L}_{SEL}$ replace $\mathscr{L}_{CLL}$, and 7) 'Contrastive Loss',

Table 5.7: Evaluation of the contrastive learning loss $\mathscr{L}_{CLL}$ for HUB on ImageNet-C with ResNet-50. We list the averaged accuracy (%). In this experiment, the symmetric entropy loss is replaced by 1) a standard entropy loss, 2) a soft likelihood ratio loss $\mathscr{L}_{SLR}$, 3) a symmetric entropy loss $\mathscr{L}_{SEL}$, and 4) a $\mathscr{L}_{SLR}$ with $\mathscr{L}_{SEL}$. The result shows that our weighted contrastive learning loss achieves the highest accuracy.

Method	Average Acc.
Baseline (ETA)	39.6
EBaR-$\mathscr{L}_{CLL}$	45.6 (+6.0%)
Entropy Loss	44.2 (+4.6%)
$\mathscr{L}_{SLR}$	44.8 (+5.2%)
$\mathscr{L}_{SEL}$	44.5 (+4.9%)
$\mathscr{L}_{SLR}$ + $\mathscr{L}_{SEL}$	44.7 (+5.1%)
Contrastive Loss	45.8 (+6.2%)
EBaR	**47.6 (+8.0%)**

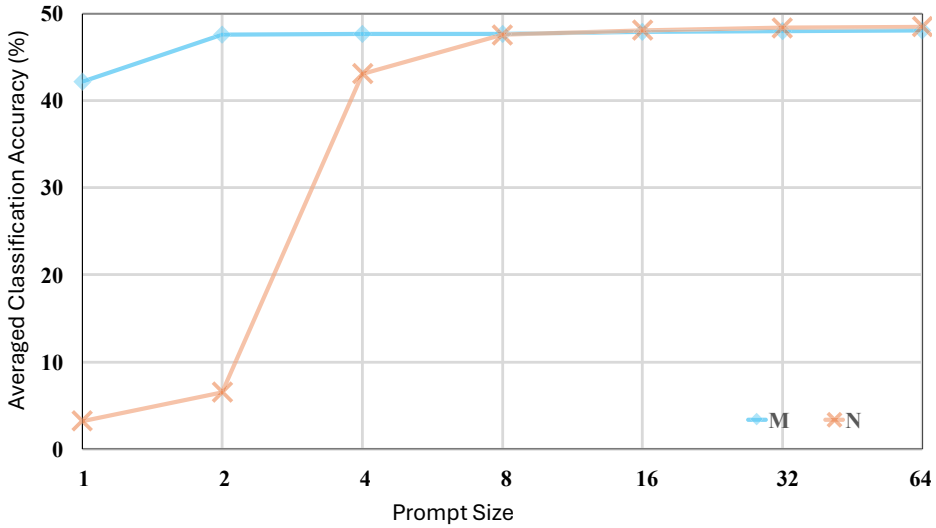


Figure 5.3: Effect of buffer sizes M and N. Averaged accuracy with different LUB size M and HUB size N on ImageNet-C using ResNet-50 is reported. The optimum setting is $M = 2, N = 8$.

where a standard contrastive learning loss is adopted without the sample weight Δ_S.

As shown in Table 5.7, without $\mathscr{L}_{CLL}$, the accuracy decreases by 2.0%. When applying the entropy loss or $\mathscr{L}_{SLR}$ or $\mathscr{L}_{SEl}$, the accuracy further decreases compared to EBaR without $\mathscr{L}_{CLL}$, indicating that the entropy loss and the soft likelihood loss are not suitable for high uncertainty samples. The reason is that the samples with high uncertainty levels generate low-quality and noisy pseudo-labels.

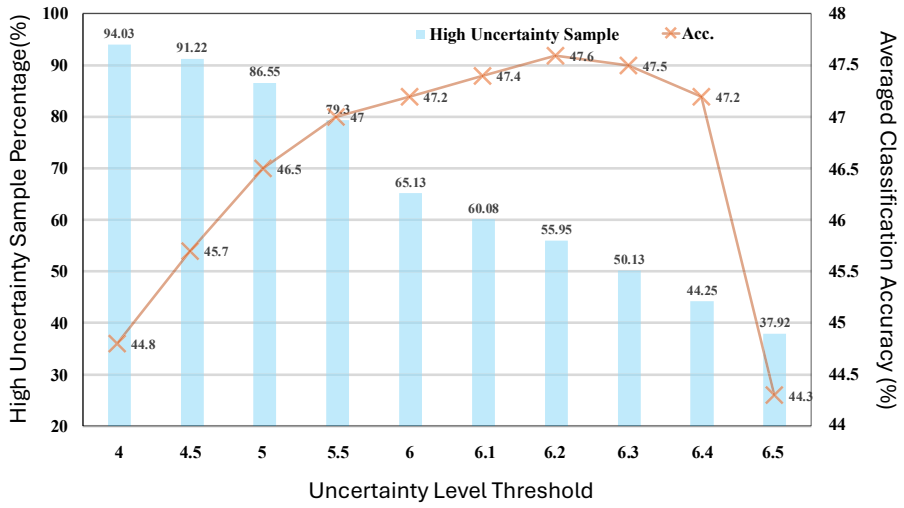


Figure 5.4: Effect of uncertainty level threshold ϵ. Experiments are conducted on ImageNet-C with ResNet-50. The percentage of samples categorised as high uncertainty and the corresponding accuracy are reported. $\epsilon = 6.2$ is the optimum.

5.3.5 Ablation Studies of Hyper-Parameters

5.3.5.1 Buffer Size

We evaluate the impact of the LUB and HUB sizes M and N (Algo.1). In the LUB experiment group, we set $N = 8$; in the HUB experiment group, we set $M = 2$. As shown in Figure 5.3, EBaR achieves optimal performance when M reaches 2, with minimal improvement when $M > 2$. A similar trend is observed for N, where performance stabilises after N reaches 8. In particular, when N is smaller than 4, EBaR collapses due to the influence of high-uncertainty samples. These samples cause a misestimation of the distribution and disrupt the updates. To balance memory efficiency and accuracy, we set $M = 2$ and $N = 8$ as the default.

5.3.5.2 Uncertainty Level Threshold

We evaluate the impact of the uncertainty level threshold ϵ (Algo.1). The result is shown in Figure 5.4. As ϵ increases, the proportion of samples classified as high uncertainty gradually decreases, drastically dropping when $\epsilon \geq 6$. The high- and low-uncertainty samples are balanced when $\epsilon \in (6.1, 6.4)$. When most samples are categorised as high uncertainty, LUB becomes ineffective. In contrast, HUB is ineffective when most are categorised as low uncertainty. EBaR achieves the highest accuracy $\epsilon = 6.2$, where samples from both levels are balanced. Thus, we set $\epsilon = 6.2$ as the default.

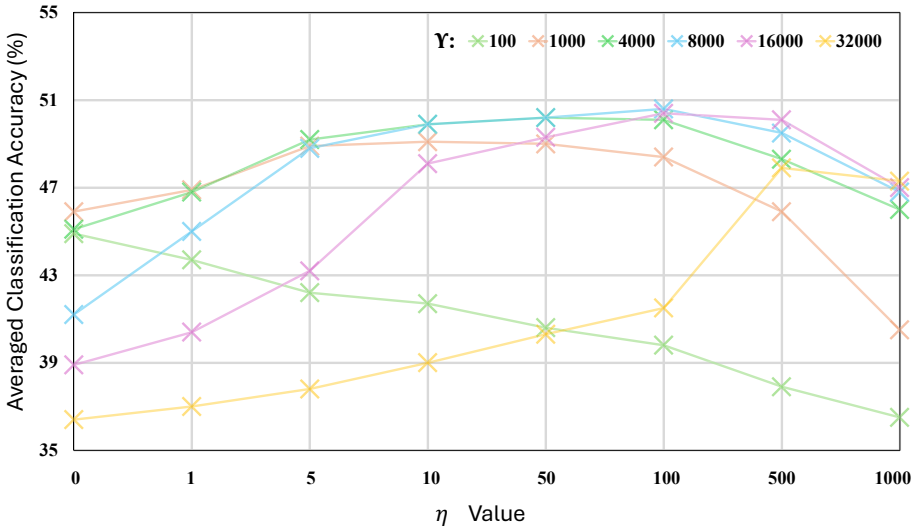


Figure 5.5: Effect of the parameters that control the reset frequency: η, and Υ (Algo.2). We fix $\mu = 1$. The experiment is conducted on the CCC with ResNet-50. Averaged accuracy is reported. The result shows that $\eta = 100, \Upsilon = 8000$ is optimum.

5.3.5.3 Resetting Frequency

We evaluate the influence of the parameters that control the reset frequency, namely μ, η, and υ (Algo.2). We fix $\mu = 1$ in the experiments and adjust the other two parameters. The reason for this is that the optimal Υ and η are proportional to μ. That being said, setting $\mu = 1, \Upsilon = 8000, \eta = 100$ is equivalent to $\mu = 2, \Upsilon = 16000, \eta = 200$. Hence, we can set $\mu = 1$ and find the optimal Υ and η.

As shown in Figure 5.5, when Υ increases, the optimal η increases. When Υ is small while η is large, the model is reset excessively frequently, harming the adaptation. Conversely, when Υ is large while η is small, the model is not reset in time, leading to catastrophic forgetting. $\eta = 100, \Upsilon = 8000$ achieve the highest accuracy and are set as the default.

5.3.5.4 Learning Rate

We evaluate the influence of the learning rate during the test-time adaptation. The result is shown in Figure 5.6. When the learning rate is greater than 1×10^{-2}, the performance declines as high-uncertainty samples severely disturb the adaptation. As the learning rate decreases, classification accuracy increases and then remains stable when the learning rate is smaller than 2.5×10^{-4}. When the learning rate is 1×10^{-4}, the

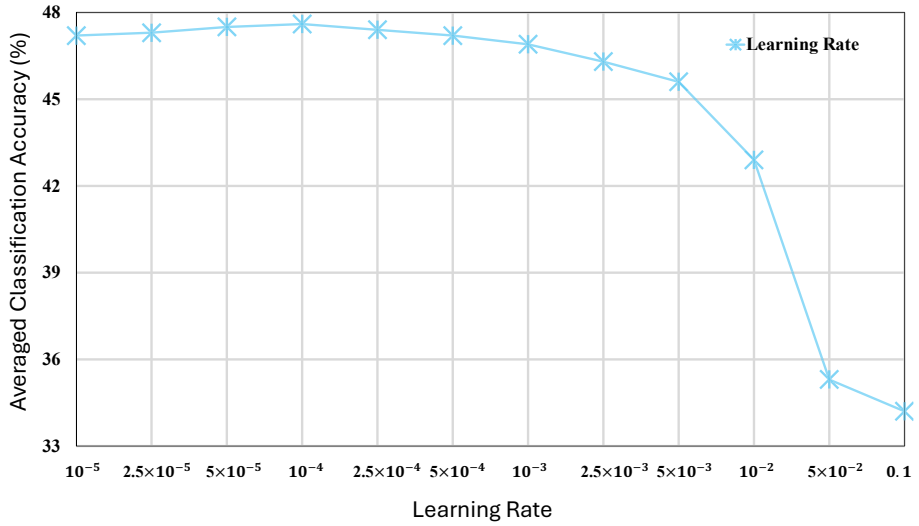


Figure 5.6: Effect of learning rate. Averaged accuracy of EBaR with different learning rates on ImageNet-C using ResNet-50. The optimum learning rate setting is 1×10^{-4}.

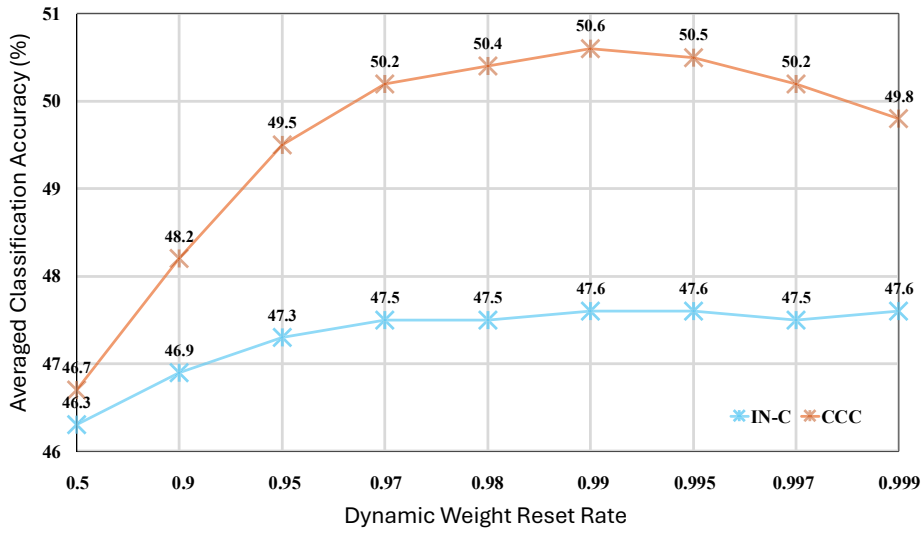


Figure 5.7: Effect of dynamic weight reset rate α. Averaged accuracy of EBaR with different α on ImageNet-C and CCC using ResNet-50. The optimum setting is $\alpha = 0.99$.

model achieves the highest accuracy of 47.6%. As a result, we set the default learning rate as 1×10^{-4}.

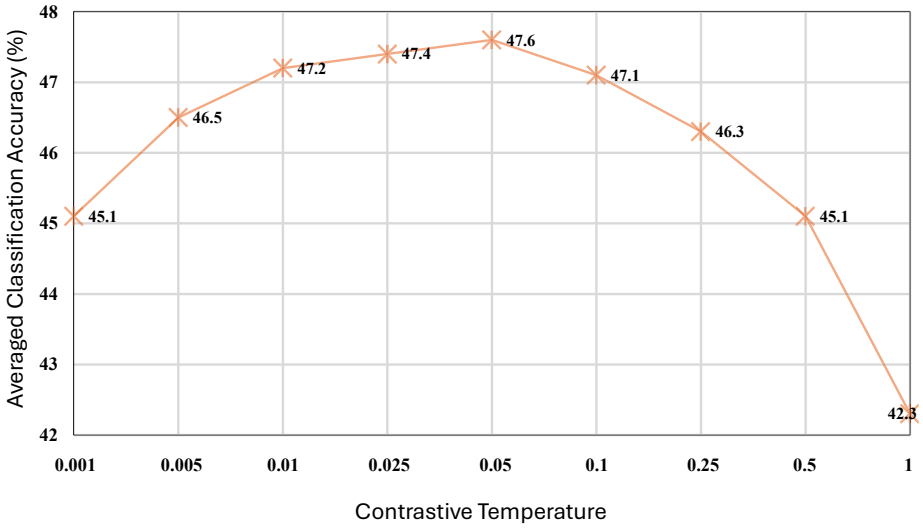


Figure 5.8: Effect of temperature τ in the contrastive learning loss $\mathscr{L}_{CLL}$. The averaged accuracy of EBaR with different learning rates on ImageNet-C using ResNet-50 is reported. The optimum learning rate setting is $\tau = 0.05$.

5.3.5.5 Dynamic Weight Reset Rate α

We evaluate the influence of the dynamic weight reset rate α during test-time adaptation. The result is shown in Figure 5.7. The accuracy for the ImageNet-C dataset increases and remains stable as α increases. As the size of the ImageNet-C dataset is around 7.5×10^5, the influence of the reset rate α is limited. On the other hand, the influence of $alpha$ is more significant in the CCC dataset with 7.5×10^6 samples. This phenomenon occurs because CCC requires the model to have a stronger ability to remember the source knowledge at extended time steps. The accuracy increases drastically before $alpha$ reaches 0.97. Then, the accuracy remains stable and slightly decreases after 0.99. EBaR achieves the highest accuracy on both ImageNet-C and CCC with $alpha = 0.99$, which is set as the default.

5.3.5.6 Contrastive Temperature

We evaluate the influence of the contrastive temperature τ in $\mathscr{L}_{CLL}$. As shown in Figure 5.8, when the temperature is above 0.5, $\mathscr{L}_{CLL}$ is ineffective, leading to collapse of EBaR and a significant decrease in the accuracy. The accuracy reaches the peak when $\tau = 0.05$. After that, the accuracy decreases as tau exceeds 0.05. As a result, we set $\tau = 0.05$ as the default.

Table 5.8: Computational efficiency of EBaR. Experiments are conducted on ImageNet-C with a ResNet-50 backbone. 'W32' represents a moving window [117, 128] of size 32. The experiments are conducted on a single RTX A5500 GPU. EBaR yields a higher accuracy boost with less extra computational overhead.

Method	Memory Cost	Time Per Prediction	Average Acc.
ETA	0.93G	3.4 ms	4.1
ETA+W32	6.22G	2.1 ms	39.6
ROID	1.13G	41.6 ms	5.2
ROID+W32	6.42G	5.3 ms	43.7
EBaR	3.72G	4.1 ms	**47.6**

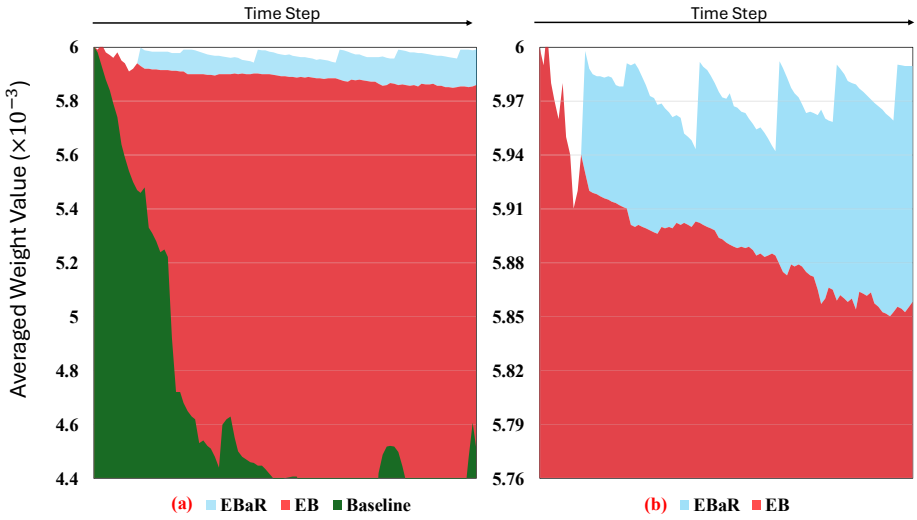


Figure 5.9: The adaptation stability of EBaR. EBaR is compared with EBaR with Efficient Buffer Only (EB), and ETA with window size 32 (Baseline) on the CCC dataset across 3×10^4 steps with ResNet-50. The averaged weight value ($\times 10^{-3}$) of a normalisation layer is depicted. The result shows that EB significantly enhances adaptation stability while elastic resetting maintains stability over the long run.

5.3.6 Observations

5.3.6.1 Computational Efficiency

We evaluate the computational efficiency of EBaR. As shown in Table 5.8, we compare the computational overhead of ETA, ROID, and EBaR on a single RTX A5500 GPU. Without a moving window, ETA and ROID both collapse. Added a moving window of size 32, ETA and ROID converge. Still, EBaR achieves a higher classification accuracy, requiring only 60% of the memory overhead.

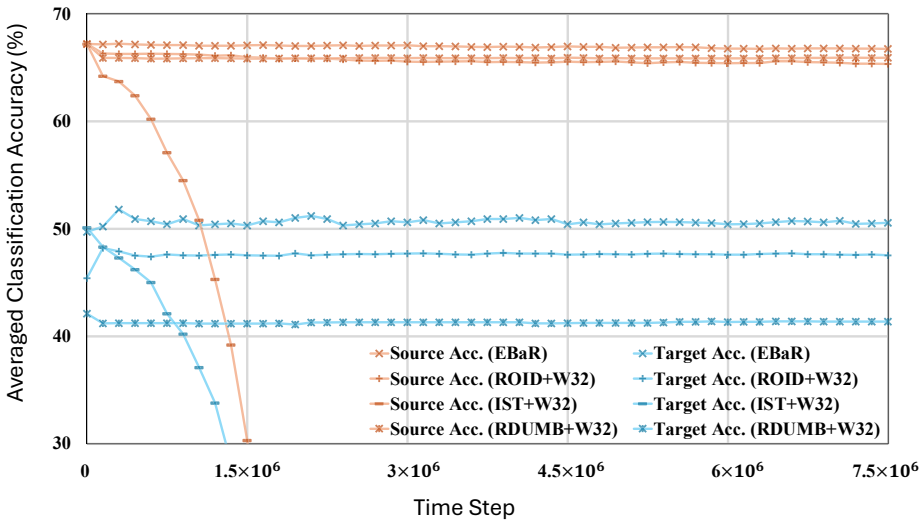


Figure 5.10: Anti-forgetting ability of EBaR. EBaR is compared with IST, ROID, and RDUMB (all three of them are with a moving window of size 32) on ImageNet (Source) and CCC (Target) using ResNet-50. The averaged accuracy per 1.5×10^5 steps across 7.5×10^6 steps is reported. The result demonstrates that EBaR effectively prevents catastrophic forgetting while retaining useful target knowledge.

Furthermore, EBaR improves accuracy by +8% with only 1.9 ms of additional time per prediction compared to the ETA baseline. In comparison, ROID achieves lower accuracy while incurring 3.2 ms of extra time per prediction. Interestingly, we observe that the moving window reduces the prediction time by decreasing the frequency of model updates.

5.3.6.2 Stable Adaptation

We examine the ability of EBaR to enhance adaptation stability by analysing the fluctuation in parameter values of normalisation layers over time. Figure 5.9 depicts the averaged weight values of a batch normalisation layer in the ResNet-50 backbone. The baseline model (ETA with window size 32) suffers from domain shift perturbations, resulting in severe fluctuations in parameter values. The efficient buffer increases robustness to perturbations and reduces fluctuations significantly. Furthermore, elastic resetting maintains adaptation stability in the long run.

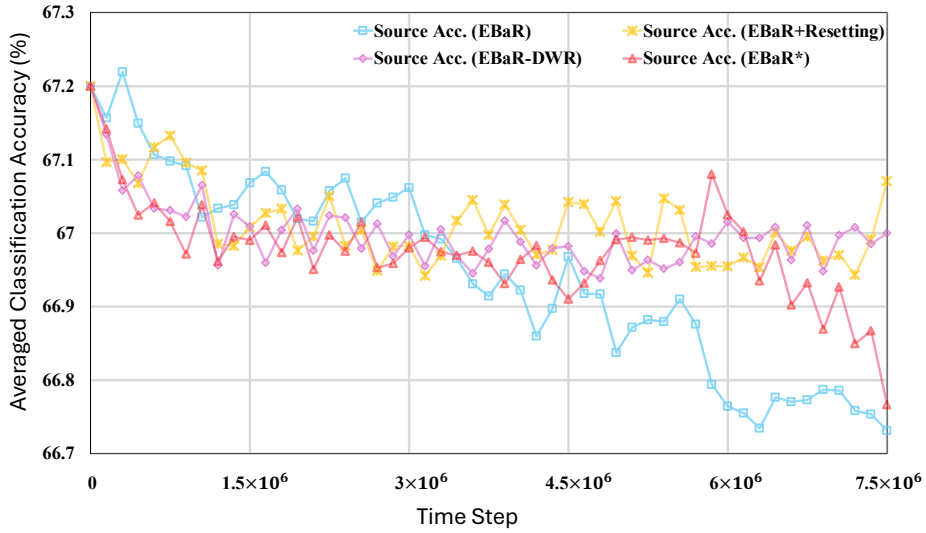


Figure 5.11: Observation of anti-forgetting ability for different resetting strategies. EBaR is compared with 1) 'EBaR+Resetting', which denotes EBaR that adopts a standard model resetting strategy where the whole model is reset to the source model every 2000 iterations [139], 2) 'EBaR-DWR', which denotes EBaR that replaces the dynamic weight reset with standard reset where the parameters are set back to the initialized values, and 3) 'EBAR*', which denotes EBaR with only dynamic weight reset and all the parameters is reset every 2000 iterations. We report the source (ImageNet) accuracy of the model (Figure 5.11) and the target (CCC) accuracy of the model (Figure 5.12). The averaged accuracy per 1.5×10^5 steps across 7.5×10^6 steps with ResNet-50 backbone is reported. The result in this figure shows that our strategy is as strong as the standard reset regarding preventing catastrophic forgetting.

5.3.6.3 Anti-forgetting

We compare the ability of EBaR to prevent catastrophic forgetting with IST, ROID, and RDUMB (three counterparts have an additional moving window of size 32), as shown in Figure 5.10, the average accuracies per 1.5×10^5 time steps for CCC (denoted as Target) and original ImageNet (denoted as Source) are reported across 7.5×10^6 time steps. EBaR achieves the highest average source accuracy (66.9%) and target accuracy (50.6%). The result demonstrates that EBaR can effectively prevent catastrophic forgetting while retaining more useful target knowledge.

5.3.6.4 Different Resetting Strategy

We compare elastic resetting with other resetting strategies. As shown in Fig.5.11 and Fig.5.12, EBaR with elastic resetting is compared with: 1) 'EBaR+Resetting', which

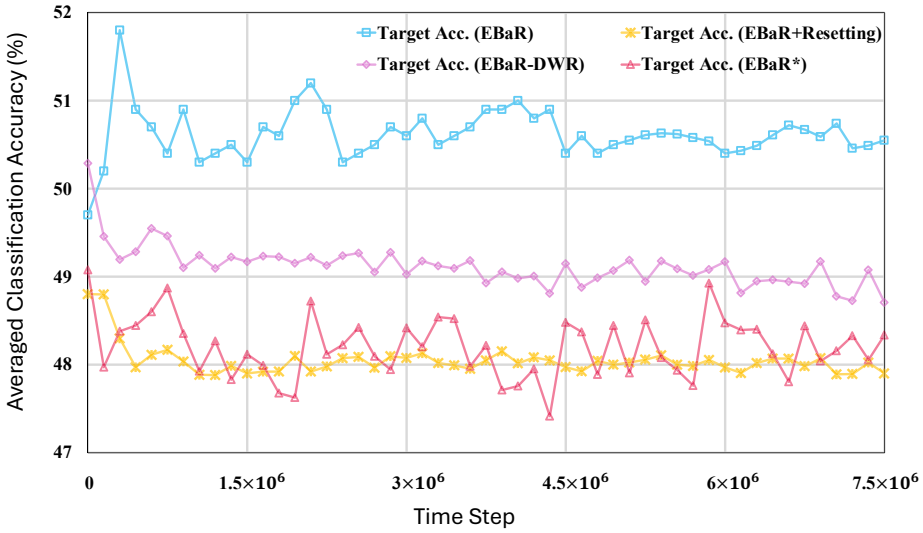


Figure 5.12: Observation of target adaptation for different resetting strategies. The averaged target (CCC) accuracy of the models is illustrated. The average accuracy per 1.5×10^5 steps across 7.5×10^6 steps with ResNet-50 backbone is reported. The result demonstrates that our elastic resetting strategy effectively prevents forgetting and allows the model to extract and retain more target knowledge. As a result, our method boosts the target accuracy significantly compared to other resetting strategies.

denotes EBaR that adopts a standard model resetting strategy where the whole model is reset to the source model every 2000 iterations [139], 2) 'EBaR-DWR', which denotes EBaR that replaces the dynamic weight reset with standard reset where the parameters are set back to the initialized values, and 3) 'EBAR*', which denotes EBaR with only dynamic weight reset and all the parameters is reset every 2000 iterations. The experiment is conducted on the CCC dataset with the ResNet-50 backbone. Fig.5.11 and Fig.5.12 illustrate the source (ImageNet) and the target (CCC) accuracy, respectively. Fig.5.11 shows that EBaR with elastic resetting achieves a source accuracy comparable to the standard hard reset, where the whole model is forced to reset to the initial source model every 2000 iterations [139]. The result shown in Fig.5.12 demonstrates the superiority of EBaR on the target dataset, indicating that the elastic resetting strategy allows the model to retain more useful target knowledge rather than sacrifice it.

5.3.6.5 Under CoTTA Setting

We compare EBaR with recent CoTTA methods on the CoTTA task under different batch size settings. In this experiment, no moving window is applied to the counterparts. As

Table 5.9: EBaR against other CoTTA counterparts under the standard CoTTA setting on ImageNet-C using ResNet-50 backbone. Averaged accuracy is reported. EBaR demonstrates more substantial advantages under small-batch settings while achieving comparable performance under large-batch settings.

Batch Size	ETA	ViDA	IST	ROID	BDG	EBaR
4	19.9	20.1	33.1	**40.5**	40.4	**47.9**
8	22.1	30.4	35.4	42.8	**42.9**	**48.2**
16	31.8	36.9	36.8	43.5	**43.6**	**48.6**
32	40.1	44.3	37.6	44.2	**44.5**	**48.8**
64	40.2	45.7	38.1	48.1	**48.9**	48.8

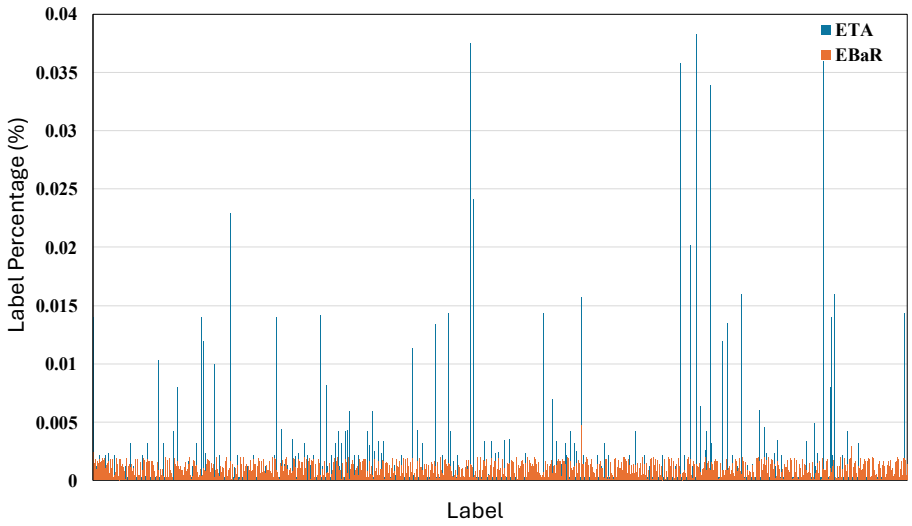


Figure 5.13: The predicted label distribution comparison under S-CoTTA. We compare the distribution of the predicted label between EBaR and ETA (with a moving window of size 16) on CCC with ResNet-50 across the first 1×10^6 time steps. The result shows that EBaR presents more balanced predictions without further restraint.

shown in Table 5.9, EBaR demonstrates a significant advantage in the small-batch settings. Even with large-batch settings, EBaR achieves performance comparable to the SOTA methods on standard CoTTA tasks, highlighting the robustness and versatility of our approach.

5.3.6.6 Distribution of Predicted Labels

We visualise the distribution of predicted labels for 1) EBaR and 2) ETA with a moving window of size 16. As shown in Figure 5.13, EBaR achieves a more balanced label prediction without further restraint. The insight is that, as noted by [149], the high

Table 5.10: Influence of different normalisation layers. Standard batch normalisation and group normalisation are compared. '+GN' denotes that the corresponding ResNet-50 model uses group normalisation and is pre-trained on ImageNet. For other groups, the model adopts a standard batch normalisation layer. The result shows that group normalisation improves the model performance when no moving window is applied. The reason is that the standard batch normalisation cannot estimate the distribution statistics based on a single sample at a time [117, 149].

Method	Average Acc.
ETA	4.1
ETA+W32	39.6
ETA+GN	37.1
ETA+GN+W32	39.4
ROID+W32	43.7
ROID+GN	42.9
EBaR	47.6
EBaR+GN	**48.1**

uncertainty samples during S-CoTTA provide highly noisy pseudo-labels and disturb the model, especially at the beginning of the adaptation. These perturbations lead to uneven predictions. EBaR mitigates the effect by 1) congregating the high-uncertainty samples in a larger batch and 2) applying a weighted contrastive learning loss based on the uncertainty level of the sample. In this way, without being disturbed by the samples with high uncertainty, EBaR achieves balanced prediction without further constraints like proposed in [117, 202].

5.3.6.7 Normalization Layer Discussion

We compare different selections of normalisation layers for S-CoTTA. As shown in Table 5.10, standard batch normalisation and group normalisation (denoted as '+GN') are compared. For the group normalisation, we adopt the ResNet-50 variant with group normalisation. The result shows that group normalisation improves the model performance when no moving window is applied. The reason is that the standard batch normalisation cannot accurately estimate distribution statistics based on a single sample at a time during self-supervised adaptation [117, 149]. Conversely, group normalisation layers estimate the distribution using sample statistics from multiple batches, avoiding a severely biased estimation. EBaR significantly improves the classification accuracy of models that adopt different normalisation layers, showing its robustness.

Table 5.11: The effectiveness of EBaR on the zero-shot model CLIP is demonstrated. We apply the proposed EBaR strategy to a recent CLIP-based test-time adaptation method, CLIPARTT [26], under the Single-Sample Continual Test-Time Adaptation (S-CoTTA) setting. The results show that EBaR significantly improves the accuracy of the CLIP model, using different backbones, across multiple benchmark datasets.

Method	Backbone	Dataset	Average Acc.
CLIPARTT	ResNet-50	CIFAR-10-C	68.1
CLIPARTT+EBaR	ResNet-50	CIFAR-10-C	72.9 **(+4.8 %)**
CLIPARTT	ViT-B-16	CIFAR-100-C	37.7
CLIPARTT+EBaR	ViT-B-16	CIFAR-100-C	41.1 **(+3.4 %)**
CLIPARTT	ViT-B-16	ImageNet-C	23.8
CLIPARTT+EBaR	ViT-B-16	ImageNet-C	27.9 **(+4.1 %)**

5.3.6.8 EBaR on Zero-Shot Models

Zero-shot classification with vision-language models, such as CLIP [142], has shown remarkable potential for addressing zero-shot image classification tasks. By leveraging a joint embedding space for images and text, these models can effectively alleviate the label shift problem without requiring task-specific fine-tuning. Nevertheless, despite their strong generalisation capabilities, CLIP and similar models still experience notable performance degradation when operating under continual domain shift conditions [126].

In this section, we investigate the integration of the proposed EBaR strategy with the CLIP model in conjunction with CLIPARTT [26], a recent method designed for test-time domain adaptation in CLIP. We evaluate CLIPARTT under the Single-Sample Continual Test-Time Adaptation (S-CoTTA) setting, both with and without EBaR. The results, presented in Table 5.11, show that EBaR substantially improves the performance of CLIP under the S-CoTTA zero-shot setting. These findings demonstrate that EBaR is a generalisable approach that can be applied to zero-shot models, enabling the joint handling of both label shift and continual domain shift.

CHAPTER 6

FUTURE WORK

6.1 Continual Domain Shift with Different Modalities: Multi-Modality and Interdisciplinary CoTTA

Recent CoTTA research has begun to focus more on multi-modality models, including vision-language models [37, 75], vision-audio models [16], and video models [42]. Addressing CoTTA in the context of modality-specific characteristics and enabling cross-modality solutions is a promising future direction.

Another promising research avenue is combining CoTTA with domain-specific or interdisciplinary knowledge. General-purpose CoTTA methods may be insufficient for handling the unique challenges of specific application scenarios. For instance, in medical imaging, different modalities such as X-rays, CT scans, and ultrasound images exhibit distinct noise patterns and perturbations. Rather than developing a universal solution, focusing on a specific modality and designing adaptation strategies tailored to its properties can be highly valuable. This approach can be extended to other interdisciplinary domains, such as remote sensing, 3D point clouds, and anomaly detection, where domain knowledge plays a crucial role in robust adaptation.

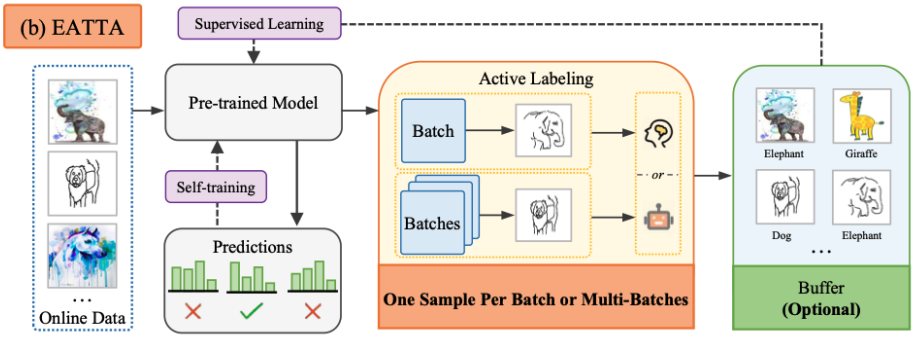


Figure 6.1: The concept of active continual test-time adaptation, depicted in [171]. Samples are actively selected for human labelling at the test time. The target is to achieve higher accuracy with fewer chosen samples to be labelled.

6.2 Continual Domain Shift with Wider Domain Gaps: Active CoTTA

Recent CoTTA methods primarily focus on addressing corruption-based domain gaps, such as variations in lighting conditions, Gaussian noise, and contrast changes. While these approaches have shown effectiveness in handling these distribution shifts, real-world scenarios often involve far more complex types of domain shift. Examples include transitions from simulation-generated images to real-world photographs, from paintings to real-world images, or from sketch drawings to fully colored paintings. Under such circumstances, the domain discrepancies are considerably more intricate and, to date, have not been effectively addressed by existing CoTTA methods.

In response to this limitation, recent work [171] proposes an effective active labelling scheme for CoTTA (Active CoTTA) that aims to handle more challenging and complex domain gaps. As illustrated in Figure 6.1, the objective is to select a minimal number of samples within each mini-batch for manual annotation by human experts. The model can then leverage these selectively labelled samples to acquire target domain-specific knowledge and recalibrate itself, thereby mitigating the accumulation of prediction errors over time. This emerging research direction demonstrates strong potential for addressing more intricate domain shift scenarios in CoTTA and is likely to play an important role in future developments in the field.

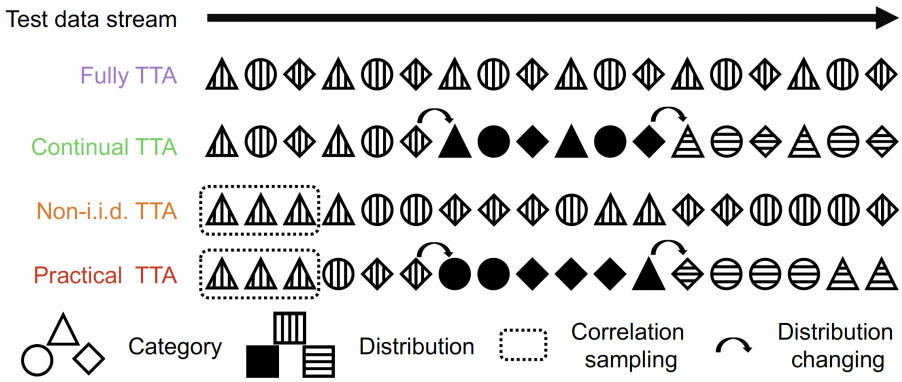


Figure 6.2: The concept of practical test-time adaptation [202]. Practical test-time adaptation is targeted at the non-i.i.d problem in CoTTA caused by temporal label correlation, which is common in real-world scenarios.

6.3 Continual Domain Shift with Label Shift

6.3.1 Non-i.i.d. CoTTA

Non-i.i.d. Continual Test-Time Adaptation (Non-i.i.d. CoTTA) refers to the situation in CoTTA where target labels are not independently and identically distributed in the test-time data stream. Non-i.i.d. CoTTA jointly considers the label shift and the continual domain shift problem.

For example, [202] proposed a practical test-time adaptation setting, which is a subclass of Non-i.i.d. CoTTA. As illustrated in Figure 6.2, practical TTA corresponds to a CoTTA scenario where the target labels exhibit temporal correlation across nearby time steps. This setting aligns well with many real-world cases where target labels in a scene are inherently correlated. Under this setting, the distribution of the target ground truth label changes over time, leading to severe label shift. In [202], a class-balanced memory bank is proposed to deal with the label shift with continuous domain shift. Though effective, the memory bank introduces large memory overhead. A future direction is to solve this issue with less computational and memory cost.

Another example is the universal test-time adaptation, which is proposed by [117] to address domain shifts under more generalised and diverse conditions, extending beyond standard CoTTA. In this setting, the model is expected to handle different scenarios, including CoTTA and mixed-domain shifts, under non-i.i.d. label distributions. Consequently, the proposed method must exhibit a high degree of generalizability and robustness to operate effectively under such heterogeneous test-time conditions.

6.3.2 Zero-Shot CoTTA

Zero-shot classification with vision-language models, such as CLIP [142], has demonstrated remarkable potential in addressing zero-shot image classification tasks. By leveraging the joint embedding space of images and text, these models can effectively mitigate the label shift problem without requiring task-specific fine-tuning. However, despite their strong generalisation capabilities, CLIP and similar models still suffer from performance degradation when deployed under continual domain shift conditions [126]. This highlights the importance and promise of developing methods that address continual domain shift in the context of zero-shot learning.

Recent works [26, 63, 177] have begun exploring test-time adaptation techniques for CLIP, aiming to enhance its robustness in dynamically changing environments. Building upon these advancements, our future work may extend such approaches to tackle more complex continual test-time adaptation scenarios (Zero-Shot CoTTA). In particular, we aim to jointly address both the label shift and continual domain shift problems, enabling zero-shot models to maintain high performance across a wide range of real-world conditions.

CHAPTER 7

CONCLUSION

In this thesis, cross-domain image classification is improved under complex real-world scenarios, explicitly addressing the challenges posed by label shift and continual domain shift. For label shift, the focus is placed on Cross-Domain Few-Shot Image Classification (CDFSIC). To tackle this problem, a novel method named Prompt-to-Disentangle (ProD) is proposed, which effectively separates source and target knowledge by leveraging a prompt-tuning mechanism. This disentanglement allows ProD to integrate the advantages of both domain generalisation and domain adaptation. ProD achieves SOTA performance across multiple benchmark datasets, and further experiments confirm the effectiveness of the proposed knowledge disentanglement scheme. For continual domain shift, the thesis primarily addresses the Continual Test-Time Adaptation (CoTTA) task. Inspired by the knowledge disentanglement strategy and findings from the CDFSIC study, a new approach, named the Source and Target Disentangle Transformer (SoTa-DiT), is introduced. SoTa-DiT explicitly disentangles source and target knowledge, enabling both the preservation of source knowledge and the extraction of target-specific features. This design achieves SOTA performance on multiple benchmarks. Additional observations verify that the model indeed learns disentangled knowledge representations, and that this disentanglement effectively addresses the dual requirements of preventing catastrophic forgetting and enabling efficient target adaptation in the CoTTA setting. Then, based on the observation that existing CoTTA methods are unstable in small-batch scenarios, a new task named Single-Sample CoTTA is explored, where the test-time batch size is set to one. An Effective Buffer and Resetting (EBaR) strategy is proposed to

improve adaptation stability for the task. Experimental results demonstrate that EBaR stabilises adaptation under the single-sample setting and prevents forgetting. Moreover, EBaR is applied to zero-shot models, enabling them to handle label shift and continual domain shift jointly. Finally, several promising future research directions are discussed, including non-i.i.d. CoTTA, universal test-time adaptation, CoTTA with multi-modality and interdisciplinary applications, active CoTTA, and CoTTA for zero-shot models.

APPENDIX

BIBLIOGRAPHY

[1] A. AFRASIYABI, H. LAROCHELLE, J.-F. LALONDE, AND C. GAGNÉ, *Matching feature sets for few-shot image classification*, in Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, 2022, pp. 9014–9024.

[2] E. ALBERTI, A. TAVERA, C. MASONE, AND B. CAPUTO, *Idda: A large-scale multi-domain dataset for autonomous driving*, IEEE Robotics and Automation Letters, 5 (2020), pp. 5526–5533.

[3] S. AO, X. LI, AND C. LING, *Fast generalized distillation for semi-supervised domain adaptation*, in Proceedings of the AAAI Conference on Artificial Intelligence, vol. 31, 2017.

[4] B. ATHIWARATKUN, M. FINZI, P. IZMAILOV, AND A. G. WILSON, *There are many consistent explanations of unlabeled data: Why you should average*, arXiv preprint arXiv:1806.05594, (2018).

[5] K. AZIZZADENESHELI, A. LIU, F. YANG, AND A. ANANDKUMAR, *Regularized learning for domain adaptation under label shifts*, arXiv preprint arXiv:1903.09734, (2019).

[6] A. BARTLER, A. BÜHLER, F. WIEWEL, M. DÖBLER, AND B. YANG, *Mt3: Meta test-time training for self-supervised test-time adaption*, in International Conference on Artificial Intelligence and Statistics, PMLR, 2022, pp. 3080–3090.

[7] D. BERTHELOT, N. CARLINI, E. D. CUBUK, A. KURAKIN, K. SOHN, H. ZHANG, AND C. RAFFEL, *Remixmatch: Semi-supervised learning with distribution alignment and augmentation anchoring*, arXiv preprint arXiv:1911.09785, (2019).

[8] D. BERTHELOT, R. ROELOFS, K. SOHN, N. CARLINI, AND A. KURAKIN, *Adamatch: A unified approach to semi-supervised learning and domain adaptation*, arXiv preprint arXiv:2106.04732, (2021).

[9] G. BLANCHARD, G. LEE, AND C. SCOTT, *Generalizing from several related classification tasks to a new unlabeled sample*, Advances in neural information processing systems, 24 (2011).

[10] D. BRAHMA AND P. RAI, *A probabilistic framework for lifelong test-time adaptation*, in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2023, pp. 3582–3591.

[11] S. BRANSON, C. WAH, F. SCHROFF, B. BABENKO, P. WELINDER, P. PERONA, AND S. BELONGIE, *Visual recognition with humans in the loop*, in European Conference on Computer Vision, Springer, 2010, pp. 438–451.

[12] L. BREIMAN, *Random forests*, Machine learning, 45 (2001), pp. 5–32.

[13] T. B. BROWN, B. MANN, N. RYDER, M. SUBBIAH, J. KAPLAN, P. DHARIWAL, A. NEELAKANTAN, P. SHYAM, G. SASTRY, A. ASKELL, S. AGARWAL, A. HERBERT-VOSS, G. KRUEGER, T. HENIGHAN, R. CHILD, A. RAMESH, D. M. ZIEGLER, J. WU, C. WINTER, C. HESSE, M. CHEN, E. SIGLER, M. LITWIN, S. GRAY, B. CHESS, J. CLARK, C. BERNER, S. MCCANDLISH, A. RADFORD, I. SUTSKEVER, AND D. AMODEI, *Language models are few-shot learners*, 2020.

[14] Z. CAI, A. RAVICHANDRAN, P. FAVARO, M. WANG, D. MODOLO, R. BHOTIKA, Z. TU, AND S. SOATTO, *Semi-supervised vision transformers at scale*, Advances in Neural Information Processing Systems, 35 (2022), pp. 25697–25710.

[15] Z. CAI, A. RAVICHANDRAN, S. MAJI, C. FOWLKES, Z. TU, AND S. SOATTO, *Exponential moving average normalization for self-supervised and semi-supervised learning*, in Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, 2021, pp. 194–203.

[16] H. CAO, Y. XU, J. YANG, P. YIN, S. YUAN, AND L. XIE, *Multi-modal continual test-time adaptation for 3d semantic segmentation*, in Proceedings of the IEEE/CVF International Conference on Computer Vision, 2023, pp. 18809–18819.

[17] G. CHAKRABARTY, M. SREENIVAS, AND S. BISWAS, *Santa: Source anchoring network and target alignment for continual test time adaptation*, Transactions on Machine Learning Research, (2023).

[18] O. CHAPELLE AND A. ZIEN, *Semi-supervised classification by low density separation*, in International workshop on artificial intelligence and statistics, PMLR, 2005, pp. 57–64.

[19] D. CHEN, D. WANG, T. DARRELL, AND S. EBRAHIMI, *Contrastive test-time adaptation*, in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2022, pp. 295–305.

[20] Q. CHEN, Z. CHEN, AND W. LUO, *Feature transformation for cross-domain few-shot remote sensing scene classification*, in International Conference on Pattern Recognition, Springer, 2022, pp. 303–316.

[21] T. CHEN, S. KORNBLITH, M. NOROUZI, AND G. HINTON, *A simple framework for contrastive learning of visual representations*, in International conference on machine learning, PmLR, 2020, pp. 1597–1607.

[22] W. CHEN, Z. ZHANG, W. WANG, L. WANG, Z. WANG, AND T. TAN, *Cross-domain cross-set few-shot learning via learning compact and aligned representations*, in European Conference on Computer Vision, Springer, 2022, pp. 383–399.

[23] W.-Y. CHEN, Y.-C. LIU, Z. KIRA, Y.-C. F. WANG, AND J.-B. HUANG, *A closer look at few-shot classification*, in International Conference on Learning Representations, 2019.

[24] X. CHEN, C.-J. HSIEH, AND B. GONG, *When vision transformers outperform resnets without pretraining or strong data augmentations*, arXiv preprint arXiv:2106.01548, (2021).

[25] Y. CHEN, X. ZHU, W. LI, AND S. GONG, *Semi-supervised learning under class distribution mismatch*, in Proceedings of the AAAI conference on artificial intelligence, vol. 34, 2020, pp. 3569–3576.

[26] Z. CHI, L. GU, H. LIU, Z. WANG, Y. WU, Y. WANG, AND K. N. PLATANIOTIS, *Learning to adapt frozen clip for few-shot test-time domain adaptation*, arXiv preprint arXiv:2506.17307, (2025).

[27] K. CHOWDHARY, *Natural language processing*, Fundamentals of artificial intelligence, (2020), pp. 603–649.

[28] S. CICEK AND S. SOATTO, *Unsupervised domain adaptation via regularized conditional alignment*, in Proceedings of the IEEE/CVF international conference on computer vision, 2019, pp. 1416–1425.

[29] N. C. CODELLA, D. GUTMAN, M. E. CELEBI, B. HELBA, M. A. MARCHETTI, S. W. DUSZA, A. KALLOO, K. LIOPYRIS, N. MISHRA, H. KITTLER, ET AL., *Skin lesion analysis toward melanoma detection: A challenge at the 2017 international symposium on biomedical imaging (isbi), hosted by the international skin imaging collaboration (isic)*, in 2018 IEEE 15th international symposium on biomedical imaging (ISBI 2018), IEEE, 2018, pp. 168–172.

[30] G. CSURKA, *Domain adaptation for visual applications: A comprehensive survey*, arXiv preprint arXiv:1702.05374, (2017).

[31] J. DAI AND S. M. BROWN, *Label bias, label shift: Fair machine learning with unreliable labels*, in NeurIPS 2020 Workshop on Consequential Decision Making in Dynamic Environments, vol. 12, 2020.

[32] D. DAS, S. YUN, AND F. PORIKLI, *ConfeSS: A framework for single source cross-domain few-shot learning*, in International Conference on Learning Representations, 2022.

[33] M. DE LANGE, R. ALJUNDI, M. MASANA, S. PARISOT, X. JIA, A. LEONARDIS, G. SLABAUGH, AND T. TUYTELAARS, *A continual learning survey: Defying forgetting in classification tasks*, IEEE transactions on pattern analysis and machine intelligence, 44 (2021), pp. 3366–3385.

[34] 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, Ieee, 2009, pp. 248–255.

[35] S. DENG, D. LIAO, X. GAO, J. ZHAO, AND K. YE, *A survey on cross-domain few-shot image classification*, in International Conference on Big Data, Springer, 2023, pp. 3–17.

[36] G. S. DHILLON, P. CHAUDHARI, A. RAVICHANDRAN, AND S. SOATTO, *A baseline for few-shot image classification*, arXiv preprint arXiv:1909.02729, (2019).

[37] M. DÖBLER, R. A. MARSDEN, T. RAICHLE, AND B. YANG, *A lost opportunity for vision-language models: a comparative study of online test-time adaptation for*

vision-language models, in European Conference on Computer Vision, Springer, 2025, pp. 117–133.

[38] M. DÖBLER, R. A. MARSDEN, AND B. YANG, *Robust mean teacher for continual and gradual test-time adaptation*, in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2023, pp. 7704–7714.

[39] A. DOSOVITSKIY, L. BEYER, A. KOLESNIKOV, D. WEISSENBORN, X. ZHAI, T. UNTERTHINER, M. DEHGHANI, M. MINDERER, G. HEIGOLD, S. GELLY, ET AL., *An image is worth 16x16 words: Transformers for image recognition at scale*, arXiv preprint arXiv:2010.11929, (2020).

[40] A. DOSOVITSKIY, L. BEYER, A. KOLESNIKOV, D. WEISSENBORN, X. ZHAI, T. UNTERTHINER, M. DEHGHANI, M. MINDERER, G. HEIGOLD, S. GELLY, J. USZKOREIT, AND N. HOULSBY, *An image is worth 16x16 words: Transformers for image recognition at scale*, 2020.

[41] L. DUAN, D. XU, AND I. W.-H. TSANG, *Domain adaptation from multiple sources: A domain-dependent regularization approach*, IEEE Transactions on neural networks and learning systems, 23 (2012), pp. 504–518.

[42] M. A.-N. I. FAHIM, M. INNAT, AND J. BOUTELLIER, *St2st: Self-supervised test-time adaptation for video action recognition*, in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2024, pp. 1057–1066.

[43] A. FARAHANI, S. VOGHOEI, K. RASHEED, AND H. R. ARABNIA, *A brief review of domain adaptation*, Advances in data science and information engineering: proceedings from ICDATA 2020 and IKE 2020, (2021), pp. 877–894.

[44] C. FINN, P. ABBEEL, AND S. LEVINE, *Model-agnostic meta-learning for fast adaptation of deep networks*, in Proceedings of the 34th International Conference on Machine Learning, D. Precup and Y. W. Teh, eds., vol. 70 of Proceedings of Machine Learning Research, PMLR, 06–11 Aug 2017, pp. 1126–1135.

[45] F. FLEURET ET AL., *Test time adaptation through perturbation robustness*, in NeurIPS 2021 Workshop on Distribution Shifts: Connecting Methods and Applications, 2021.

[46] P. FORET, A. KLEINER, H. MOBAHI, AND B. NEYSHABUR, *Sharpness-aware minimization for efficiently improving generalization*, arXiv preprint arXiv:2010.01412, (2020).

[47] C. GAN, T. YANG, AND B. GONG, *Learning attributes equals multi-source domain generalization*, in Proceedings of the IEEE conference on computer vision and pattern recognition, 2016, pp. 87–97.

[48] Y. GAN, Y. BAI, Y. LOU, X. MA, R. ZHANG, N. SHI, AND L. LUO, *Decorate the newcomers: Visual domain prompt for continual test time adaptation*, in Proceedings of the AAAI Conference on Artificial Intelligence, vol. 37, 2023, pp. 7595–7603.

[49] J. GAO, J. ZHANG, X. LIU, T. DARRELL, E. SHELHAMER, AND D. WANG, *Back to the source: Diffusion-driven adaptation to test-time corruption*, in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2023, pp. 11786–11796.

[50] S. GARG, S. BALAKRISHNAN, AND Z. LIPTON, *Domain adaptation under open set label shift*, Advances in Neural Information Processing Systems, 35 (2022), pp. 22531–22546.

[51] S. GARG, N. ERICKSON, J. SHARPNACK, A. SMOLA, S. BALAKRISHNAN, AND Z. C. LIPTON, *Rlsbench: Domain adaptation under relaxed label shift*, in International Conference on Machine Learning, PMLR, 2023, pp. 10879–10928.

[52] M. GOETZ, C. WEBER, F. BINCZYK, J. POLANSKA, R. TARNAWSKI, B. BOBEK-BILLEWICZ, U. KOETHE, J. KLEESIEK, B. STIELTJES, AND K. H. MAIER-HEIN, *Dalsa: Domain adaptation for supervised learning from sparsely annotated mr images*, IEEE transactions on medical imaging, 35 (2015), pp. 184–196.

[53] T. GONG, J. JEONG, T. KIM, Y. KIM, J. SHIN, AND S.-J. LEE, *Note: Robust continual test-time adaptation against temporal correlation*, Advances in Neural Information Processing Systems, 35 (2022), pp. 27253–27266.

[54] Y. GONG, Y. YUE, W. JI, AND G. ZHOU, *Cross-domain few-shot learning based on pseudo-siamese neural network*, Scientific Reports, 13 (2023), p. 1427.

[55] I. GOODFELLOW, Y. BENGIO, A. COURVILLE, AND Y. BENGIO, *Deep learning*, vol. 1, MIT press Cambridge, 2016.

[56] J. GOU, B. YU, S. J. MAYBANK, AND D. TAO, *Knowledge distillation: A survey*, International Journal of Computer Vision, 129 (2021), pp. 1789–1819.

[57] S. GOYAL, M. SUN, A. RAGHUNATHAN, AND J. Z. KOLTER, *Test time adaptation via conjugate pseudo-labels*, Advances in Neural Information Processing Systems, 35 (2022), pp. 6204–6218.

[58] H. GUAN AND M. LIU, *Domain adaptation for medical image analysis: a survey*, IEEE Transactions on Biomedical Engineering, 69 (2021), pp. 1173–1185.

[59] J. GUAN, M. ZHANG, AND Z. LU, *Large-scale cross-domain few-shot learning*, in Proceedings of the Asian Conference on Computer Vision, 2020.

[60] Y. GUO, N. CODELLA, L. KARLINSKY, J. V. CODELLA, J. R. SMITH, K. SAENKO, T. ROSING, AND R. FERIS, *A broader study of cross-domain few-shot learning*, in ECCV (27), 2020, pp. 124–141.

[61] Y. GUO, N. C. CODELLA, L. KARLINSKY, J. V. CODELLA, J. R. SMITH, K. SAENKO, T. ROSING, AND R. FERIS, *A broader study of cross-domain few-shot learning*, 2019.

[62] Y. GUO, N. C. CODELLA, L. KARLINSKY, J. V. CODELLA, J. R. SMITH, K. SAENKO, T. ROSING, AND R. FERIS, *A broader study of cross-domain few-shot learning*, in Computer vision–ECCV 2020: 16th European conference, glasgow, UK, August 23–28, 2020, proceedings, part XXVII 16, Springer, 2020, pp. 124–141.

[63] G. A. V. HAKIM, D. OSOWIECHI, M. NOORI, M. CHERAGHALIKHANI, A. BAHRI, M. YAZDANPANAH, I. B. AYED, AND C. DESROSIERS, *Clipartt: Adaptation of clip to new domains at test time*, in 2025 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), IEEE, 2025, pp. 7092–7101.

[64] 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, 2016, pp. 770–778.

[65] X. HE, C. LI, P. ZHANG, J. YANG, AND X. E. WANG, *Parameter-efficient model adaptation for vision transformers*, in Proceedings of the AAAI Conference on Artificial Intelligence, vol. 37, 2023, pp. 817–825.

[66] M. HEIDARI, A. ALCHIHABI, Q. EN, AND Y. GUO, *Adaptive parametric prototype learning for cross-domain few-shot classification*, AISTATS, (2024).

[67] P. HELBER, B. BISCHKE, A. DENGEL, AND D. BORTH, *Eurosat: A novel dataset and deep learning benchmark for land use and land cover classification*, IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing, 12 (2019), pp. 2217–2226.

[68] D. HENDRYCKS, S. BASART, N. MU, S. KADAVATH, F. WANG, E. DORUNDO, R. DESAI, T. ZHU, S. PARAJULI, M. GUO, ET AL., *The many faces of robustness: A critical analysis of out-of-distribution generalization*, in Proceedings of the IEEE/CVF international conference on computer vision, 2021, pp. 8340–8349.

[69] D. HENDRYCKS AND T. DIETTERICH, *Benchmarking neural network robustness to common corruptions and perturbations*, Proceedings of the International Conference on Learning Representations, (2019).

[70] J. HIRSCHBERG AND C. D. MANNING, *Advances in natural language processing*, Science, 349 (2015), pp. 261–266.

[71] L. HU, W. HE, L. ZHANG, AND H. ZHANG, *Cross-domain meta-learning under dual-adjustment mode for few-shot hyperspectral image classification*, IEEE Transactions on Geoscience and Remote Sensing, 61 (2023), pp. 1–16.

[72] Z. HU, Y. SUN, AND Y. YANG, *Switch to generalize: Domain-switch learning for cross-domain few-shot classification*, in International Conference on Learning Representations, 2021.

[73] J. HUANG AND B. KINGSBURY, *Audio-visual deep learning for noise robust speech recognition*, in 2013 IEEE international conference on acoustics, speech and signal processing, IEEE, 2013, pp. 7596–7599.

[74] X. HUANG AND S. BELONGIE, *Arbitrary style transfer in real-time with adaptive instance normalization*, in Proceedings of the IEEE international conference on computer vision, 2017, pp. 1501–1510.

[75] R. IMAM, H. GANI, M. HUZAIFA, AND K. NANDAKUMAR, *Test-time low rank adaptation via confidence maximization for zero-shot generalization of vision-language models*, in 2025 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), IEEE, 2025, pp. 5449–5459.

[76] S. IOFFE AND C. SZEGEDY, *Batch normalization: Accelerating deep network training by reducing internal covariate shift*, in International conference on machine learning, pmlr, 2015, pp. 448–456.

[77] M. JANG, S.-Y. CHUNG, AND H. W. CHUNG, *Test-time adaptation via self-training with nearest neighbor information*, arXiv preprint arXiv:2207.10792, (2022).

[78] M. JIA, L. TANG, B.-C. CHEN, C. CARDIE, S. BELONGIE, B. HARIHARAN, AND S.-N. LIM, *Visual prompt tuning*, in European Conference on Computer Vision, Springer, 2022, pp. 709–727.

[79] S. JIE AND Z.-H. DENG, *Fact: Factor-tuning for lightweight adaptation on vision transformer*, in Proceedings of the AAAI Conference on Artificial Intelligence, vol. 37, 2023, pp. 1060–1068.

[80] Y. JING, X. LIU, Y. DING, X. WANG, E. DING, M. SONG, AND S. WEN, *Dynamic instance normalization for arbitrary style transfer*, in Proceedings of the AAAI conference on artificial intelligence, vol. 34, 2020, pp. 4369–4376.

[81] A. G. KHOEE, Y. YU, AND R. FELDT, *Domain generalization through meta-learning: a survey*, Artificial Intelligence Review, 57 (2024), p. 285.

[82] A. KHOSLA, T. ZHOU, T. MALISIEWICZ, A. A. EFROS, AND A. TORRALBA, *Undoing the damage of dataset bias*, in Computer Vision – ECCV 2012, A. Fitzgibbon, S. Lazebnik, P. Perona, Y. Sato, and C. Schmid, eds., Berlin, Heidelberg, 2012, Springer Berlin Heidelberg, pp. 158–171.

[83] T. KOJIMA, Y. IWASAWA, AND Y. MATSUO, *Robustifying vision transformer without retraining from scratch using attention-based test-time adaptation*, New Generation Computing, 41 (2023), pp. 5–24.

[84] T. KOJIMA, Y. MATSUO, AND Y. IWASAWA, *Robustifying vision transformer without retraining from scratch by test-time class-conditional feature alignment*, arXiv preprint arXiv:2206.13951, (2022).

[85] W. M. KOUW AND M. LOOG, *A review of domain adaptation without target labels*, IEEE transactions on pattern analysis and machine intelligence, 43 (2019), pp. 766–785.

[86] J. KRAUSE, M. STARK, J. DENG, AND L. FEI-FEI, *3d object representations for fine-grained categorization*, in Proceedings of the IEEE international conference on computer vision workshops, 2013, pp. 554–561.

[87] A. KRIZHEVSKY, G. HINTON, ET AL., *Learning multiple layers of features from tiny images*, (2009).

[88] A. KRIZHEVSKY, I. SUTSKEVER, AND G. E. HINTON, *Imagenet classification with deep convolutional neural networks*, Advances in neural information processing systems, 25 (2012).

[89] S. LAINE AND T. AILA, *Temporal ensembling for semi-supervised learning*, arXiv preprint arXiv:1610.02242, (2016).

[90] S.-S. LEARNING, *Semi-supervised learning*, CSZ2006. html, 5 (2006), p. 1.

[91] Y. LECUN, Y. BENGIO, AND G. HINTON, *Deep learning*, nature, 521 (2015), pp. 436–444.

[92] D. LEE, J. YOON, AND S. J. HWANG, *Becotta: Input-dependent online blending of experts for continual test-time adaptation*, arXiv preprint arXiv:2402.08712, (2024).

[93] J.-H. LEE AND J.-H. CHANG, *Stationary latent weight inference for unreliable observations from online test-time adaptation*, in Forty-first International Conference on Machine Learning, 2024.

[94] X. Y. LEE, L. VIDYARATNE, M. ALAM, A. FARAHAT, D. GHOSH, T. G. DIAZ, AND C. GUPTA, *Xdnet: A few-shot meta-learning approach for cross-domain visual inspection*, in Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, 2023, pp. 4375–4384.

[95] H. LI, S. J. PAN, S. WANG, AND A. C. KOT, *Domain generalization with adversarial feature learning*, in Proceedings of the IEEE conference on computer vision and pattern recognition, 2018, pp. 5400–5409.

[96] J. LI, Z. YU, Z. DU, L. ZHU, AND H. T. SHEN, *A comprehensive survey on source-free domain adaptation*, IEEE Transactions on Pattern Analysis and Machine Intelligence, 46 (2024), pp. 5743–5762.

[97] P. LI, S. GONG, C. WANG, AND Y. FU, *Ranking distance calibration for cross-domain few-shot learning*, 2021.

[98] W. LI, L. DUAN, D. XU, AND I. W. TSANG, *Learning with augmented features for supervised and semi-supervised heterogeneous domain adaptation*, IEEE Transactions on Pattern analysis and machine intelligence, 36 (2013), pp. 1134–1148.

[99] H. LIANG, Q. ZHANG, P. DAI, AND J. LU, *Boosting the generalization capability in cross-domain few-shot learning via noise-enhanced supervised autoencoder*, 2021 IEEE/CVF International Conference on Computer Vision (ICCV), (2021), pp. 9404–9414.

[100] J. LIANG, R. HE, AND T. TAN, *A comprehensive survey on test-time adaptation under distribution shifts*, International Journal of Computer Vision, 133 (2025), pp. 31–64.

[101] J. LIANG, R. HOU, H. CHANG, B. MA, S. SHAN, AND X. CHEN, *Generalized semi-supervised learning via self-supervised feature adaptation*, Advances in Neural Information Processing Systems, 36 (2023), pp. 60791–60803.

[102] X. LIANG, Y. ZHANG, AND J. ZHANG, *Attention multisource fusion-based deep few-shot learning for hyperspectral image classification*, IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing, 14 (2021), pp. 8773–8788.

[103] B. LIU, Z. ZHAO, Z. LI, J. JIANG, Y. GUO, AND J. YE, *Feature transformation ensemble model with batch spectral regularization for cross-domain few-shot classification*, 2020.

[104] G. LIU, Z. ZHANG, AND X. FANG, *Task-adaptive multi-source representations for few-shot image recognition*, Information, 15 (2024), p. 293.

[105] J. LIU, R. XU, S. YANG, R. ZHANG, Q. ZHANG, Z. CHEN, Y. GUO, AND S. ZHANG, *Continual-mae: Adaptive distribution masked autoencoders for continual test-time adaptation*, in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2024, pp. 28653–28663.

[106] J. LIU, S. YANG, P. JIA, R. ZHANG, M. LU, Y. GUO, W. XUE, AND S. ZHANG, *Vida: Homeostatic visual domain adapter for continual test time adaptation*, arXiv preprint arXiv:2306.04344, (2023).

[107] P. LIU, W. YUAN, J. FU, Z. JIANG, H. HAYASHI, AND G. NEUBIG, *Pre-train, prompt, and predict: A systematic survey of prompting methods in natural language processing*, 2021.

[108] Q. LIU AND W. CAO, *Geometric algebra graph neural network for cross-domain few-shot classification*, Applied Intelligence, 52 (2022), pp. 12422–12435.

[109] Y. LIU, J. DENG, J. TAO, T. CHU, L. DUAN, AND W. LI, *Undoing the damage of label shift for cross-domain semantic segmentation*, in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2022, pp. 7042–7052.

[110] Y. LIU, H. ZHANG, W. ZHANG, G. LU, Q. TIAN, AND N. LING, *Few-shot image classification: Current status and research trends*, Electronics, 11 (2022), p. 1752.

[111] Y. LIU, Y. ZOU, R. LI, AND Y. LI, *Spectral decomposition and transformation for cross-domain few-shot learning*, Neural Networks, 179 (2024), p. 106536.

[112] Z. LIU, Y. LIN, Y. CAO, H. HU, Y. WEI, Z. ZHANG, S. LIN, AND B. GUO, *Swin transformer: Hierarchical vision transformer using shifted windows*, in Proceedings of the IEEE/CVF international conference on computer vision, 2021, pp. 10012–10022.

[113] K. LU, A. GROVER, P. ABBEEL, AND I. MORDATCH, *Reset-free lifelong learning with skill-space planning*, arXiv preprint arXiv:2012.03548, (2020).

[114] F. LYU, K. DU, Y. LI, H. ZHAO, Z. ZHANG, G. LIU, AND L. WANG, *Variational continual test-time adaptation*, arXiv preprint arXiv:2402.08182, (2024).

[115] J. MA, *Improved self-training for test-time adaptation*, in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2024, pp. 23701–23710.

[116] T. MA, Y. SUN, Z. YANG, AND Y. YANG, *Prod: Prompting-to-disentangle domain knowledge for cross-domain few-shot image classification*, in Proceedings of

the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2023, pp. 19754–19763.

[117] R. A. MARSDEN, M. DÖBLER, AND B. YANG, *Universal test-time adaptation through weight ensembling, diversity weighting, and prior correction*, in Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, 2024, pp. 2555–2565.

[118] A. MIKOŁAJCZYK AND M. GROCHOWSKI, *Data augmentation for improving deep learning in image classification problem*, in 2018 international interdisciplinary PhD workshop (IIPhDW), IEEE, 2018, pp. 117–122.

[119] S. MINAEE, Y. BOYKOV, F. PORIKLI, A. PLAZA, N. KEHTARNAVAZ, AND D. TERZOPOULOS, *Image segmentation using deep learning: A survey*, IEEE transactions on pattern analysis and machine intelligence, 44 (2021), pp. 3523–3542.

[120] M. J. MIRZA, J. MICOREK, H. POSSEGGER, AND H. BISCHOF, *The norm must go on: Dynamic unsupervised domain adaptation by normalization*, in Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, 2022, pp. 14765–14775.

[121] S. P. MOHANTY, D. P. HUGHES, AND M. SALATHÉ, *Using deep learning for image-based plant disease detection*, Frontiers in plant science, 7 (2016), p. 215232.

[122] C. K. MUMMADI, R. HUTMACHER, K. RAMBACH, E. LEVINKOV, T. BROX, AND J. H. METZEN, *Test-time adaptation to distribution shift by confidence maximization and input transformation*, arXiv preprint arXiv:2106.14999, (2021).

[123] H. NAM AND H.-E. KIM, *Batch-instance normalization for adaptively style-invariant neural networks*, Advances in Neural Information Processing Systems, 31 (2018).

[124] A. T. NGUYEN, T. NGUYEN-TANG, S.-N. LIM, AND P. H. TORR, *Tipi: Test time adaptation with transformation invariance*, in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2023, pp. 24162–24171.

[125] J. NI, S. YANG, R. XU, J. LIU, X. LI, W. JIAO, Z. CHEN, Y. LIU, AND S. ZHANG, *Distribution-aware continual test-time adaptation for semantic segmentation*,

in 2024 IEEE International Conference on Robotics and Automation (ICRA), IEEE, 2024, pp. 3044–3050.

[126] J. NICOLAS, F. CHIARONI, I. ZIKO, O. AHMAD, C. DESROSIERS, AND J. DOLZ, *Mop-clip: A mixture of prompt-tuned clip models for domain incremental learning*, arXiv preprint arXiv:2307.05707, (2023).

[127] F. F. NILOY, S. M. AHMED, D. S. RAYCHAUDHURI, S. OYMAK, AND A. K. ROY-CHOWDHURY, *Effective restoration of source knowledge in continual test time adaptation*, in Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, 2024, pp. 2091–2100.

[128] S. NIU, J. WU, Y. ZHANG, Y. CHEN, S. ZHENG, P. ZHAO, AND M. TAN, *Efficient test-time model adaptation without forgetting*, in International conference on machine learning, PMLR, 2022, pp. 16888–16905.

[129] S. NIU, J. WU, Y. ZHANG, Z. WEN, Y. CHEN, P. ZHAO, AND M. TAN, *Towards stable test-time adaptation in dynamic wild world*, arXiv preprint arXiv:2302.12400, (2023).

[130] M. ORBES-ARTEAINST, J. CARDOSO, L. SØRENSEN, C. IGEL, S. OURSELIN, M. MODAT, M. NIELSEN, AND A. PAI, *Knowledge distillation for semi-supervised domain adaptation*, in OR 2.0 Context-Aware Operating Theaters and Machine Learning in Clinical Neuroimaging: Second International Workshop, OR 2.0 2019, and Second International Workshop, MLCN 2019, Held in Conjunction with MICCAI 2019, Shenzhen, China, October 13 and 17, 2019, Proceedings 2, Springer, 2019, pp. 68–76.

[131] D. W. OTTER, J. R. MEDINA, AND J. K. KALITA, *A survey of the usages of deep learning for natural language processing*, IEEE transactions on neural networks and learning systems, 32 (2020), pp. 604–624.

[132] N. PAEEDEH, M. PRATAMA, M. A. MA'SUM, W. MAYER, Z. CAO, AND R. KOWLCZYK, *Cross-domain few-shot learning via adaptive transformer networks*, Knowledge-Based Systems, 288 (2024), p. 111458.

[133] J. PARK, J. KIM, H. KWON, I. YOON, AND K. SOHN, *Layer-wise auto-weighting for non-stationary test-time adaptation*, in Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, 2024, pp. 1414–1423.

[134] A. PARNAMI AND M. LEE, *Learning from few examples: A summary of approaches to few-shot learning*, arXiv preprint arXiv:2203.04291, (2022).

[135] S. PENG, W. SONG, AND M. ESTER, *Combining domain-specific meta-learners in the parameter space for cross-domain few-shot classification*, in AAAI, 2020.

[136] 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/CVF international conference on computer vision, 2019, pp. 1406–1415.

[137] L. PEREZ AND J. WANG, *The effectiveness of data augmentation in image classification using deep learning*, arXiv preprint arXiv:1712.04621, (2017).

[138] F. POURPANAH, M. ABDAR, Y. LUO, X. ZHOU, R. WANG, C. P. LIM, X.-Z. WANG, AND Q. J. WU, *A review of generalized zero-shot learning methods*, IEEE transactions on pattern analysis and machine intelligence, 45 (2022), pp. 4051–4070.

[139] O. PRESS, S. SCHNEIDER, M. KÜMMERER, AND M. BETHGE, *Rdumb: A simple approach that questions our progress in continual test-time adaptation*, Advances in Neural Information Processing Systems, 36 (2023), pp. 39915–39935.

[140] H. PURWINS, B. LI, T. VIRTANEN, J. SCHLÜTER, S.-Y. CHANG, AND T. SAINATH, *Deep learning for audio signal processing*, IEEE Journal of Selected Topics in Signal Processing, 13 (2019), pp. 206–219.

[141] J. QUIÑONERO-CANDELA, M. SUGIYAMA, A. SCHWAIGHOFER, AND N. D. LAWRENCE, *Dataset shift in machine learning*, Mit Press, 2022.

[142] A. RADFORD, J. W. KIM, C. HALLACY, A. RAMESH, G. GOH, S. AGARWAL, G. SASTRY, A. ASKELL, P. MISHKIN, J. CLARK, ET AL., *Learning transferable visual models from natural language supervision*, in International conference on machine learning, PmLR, 2021, pp. 8748–8763.

[143] I. RADOSAVOVIC, P. DOLLÁR, R. GIRSHICK, G. GKIOXARI, AND K. HE, *Data distillation: Towards omni-supervised learning*, in Proceedings of the IEEE conference on computer vision and pattern recognition, 2018, pp. 4119–4128.

[144] W. RAWAT AND Z. WANG, *Deep convolutional neural networks for image classification: A comprehensive review*, Neural computation, 29 (2017), pp. 2352–2449.

[145] M. REN, E. TRIANTAFILLOU, S. RAVI, J. SNELL, K. SWERSKY, J. B. TENENBAUM, H. LAROCHELLE, AND R. S. ZEMEL, *Meta-learning for semi-supervised few-shot classification*, arXiv preprint arXiv:1803.00676, (2018).

[146] M. ROSTAMI, S. KOLOURI, E. EATON, AND K. KIM, *Sar image classification using few-shot cross-domain transfer learning*, in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops, 2019, pp. 0–0.

[147] E. RUSAK, S. SCHNEIDER, P. V. GEHLER, O. BRINGMANN, W. BRENDEL, AND M. BETHGE, *Imagenet-d: A new challenging robustness dataset inspired by domain adaptation*, in ICML 2022 Shift Happens Workshop, 2022.

[148] O. RUSSAKOVSKY, J. DENG, H. SU, J. KRAUSE, S. SATHEESH, S. MA, Z. HUANG, A. KARPATHY, A. KHOSLA, M. BERNSTEIN, A. C. BERG, AND L. FEI-FEI, *ImageNet Large Scale Visual Recognition Challenge*, International Journal of Computer Vision (IJCV), 115 (2015), pp. 211–252.

[149] S. SETO, B.-J. THEOBALD, F. DANIELI, N. JAITLY, AND D. BUSBRIDGE, *Realm: Robust entropy adaptive loss minimization for improved single-sample test-time adaptation*, in Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, 2024, pp. 2062–2071.

[150] J. SNELL, K. SWERSKY, AND R. ZEMEL, *Prototypical networks for few-shot learning*, in Advances in Neural Information Processing Systems, I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, eds., vol. 30, Curran Associates, Inc., 2017.

[151] D. SÓJKA, S. CYGERT, B. TWARDOWSKI, AND T. TRZCIŃSKI, *Ar-tta: A simple method for real-world continual test-time adaptation*, in Proceedings of the IEEE/CVF International Conference on Computer Vision, 2023, pp. 3491–3495.

[152] J. SONG, J. LEE, I. S. KWEON, AND S. CHOI, *Ecotta: Memory-efficient continual test-time adaptation via self-distilled regularization*, in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2023, pp. 11920–11929.

[153] Y. SONG, T. WANG, P. CAI, S. K. MONDAL, AND J. P. SAHOO, *A comprehensive survey of few-shot learning: Evolution, applications, challenges, and opportunities*, ACM Computing Surveys, 55 (2023), pp. 1–40.

[154] M. SREENIVAS AND S. BISWAS, *Similar class style augmentation for efficient cross-domain few-shot learning*, in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2023, pp. 4590–4598.

[155] A. STEINER, A. KOLESNIKOV, , X. ZHAI, R. WIGHTMAN, J. USZKOREIT, AND L. BEYER, *How to train your vit? data, augmentation, and regularization in vision transformers*, arXiv preprint arXiv:2106.10270, (2021).

[156] S. SUN, H. SHI, AND Y. WU, *A survey of multi-source domain adaptation*, Information Fusion, 24 (2015), pp. 84–92.

[157] T. SUN, M. SEGU, J. POSTELS, Y. WANG, L. VAN GOOL, B. SCHIELE, F. TOMBARI, AND F. YU, *Shift: a synthetic driving dataset for continuous multi-task domain adaptation*, in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2022, pp. 21371–21382.

[158] F. SUNG, Y. YANG, L. ZHANG, T. XIANG, P. H. TORR, AND T. M. HOSPEDALES, *Learning to compare: Relation network for few-shot learning*, in 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2018, pp. 1199–1208.

[159] R. TACHET DES COMBES, H. ZHAO, Y.-X. WANG, AND G. J. GORDON, *Domain adaptation with conditional distribution matching and generalized label shift*, Advances in Neural Information Processing Systems, 33 (2020), pp. 19276–19289.

[160] M. TAN, G. CHEN, J. WU, Y. ZHANG, Y. CHEN, P. ZHAO, AND S. NIU, *Uncertainty-calibrated test-time model adaptation without forgetting*, IEEE Transactions on Pattern Analysis and Machine Intelligence, (2025).

[161] A. TARVAINEN AND H. VALPOLA, *Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results*, Advances in neural information processing systems, 30 (2017).

[162] P. TIAN AND S. XIE, *An adversarial meta-training framework for cross-domain few-shot learning*, IEEE Transactions on Multimedia, 25 (2022), pp. 6881–6891.

[163] Y. TIAN, Y. WANG, D. KRISHNAN, J. B. TENENBAUM, AND P. ISOLA, *Rethinking few-shot image classification: a good embedding is all you need?*, in European conference on computer vision, Springer, 2020, pp. 266–282.

[164] I. TOLSTIKHIN, N. HOULSBY, A. KOLESNIKOV, L. BEYER, X. ZHAI, T. UNTERTHINER, J. YUNG, A. STEINER, D. KEYSERS, J. USZKOREIT, M. LUCIC, AND A. DOSOVITSKIY, *Mlp-mixer: An all-mlp architecture for vision*, arXiv preprint arXiv:2105.01601, (2021).

[165] D. TOMAR, G. VRAY, B. BOZORGTABAR, AND J.-P. THIRAN, *Tesla: Test-time self-learning with automatic adversarial augmentation*, in Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, 2023, pp. 20341–20350.

[166] H.-Y. TSENG, H.-Y. LEE, J.-B. HUANG, AND M.-H. YANG, *Cross-domain few-shot classification via learned feature-wise transformation*, 2020.

[167] G. VAN HORN, O. MAC AODHA, Y. SONG, Y. CUI, C. SUN, A. SHEPARD, H. ADAM, P. PERONA, AND S. BELONGIE, *The inaturalist species classification and detection dataset*, in Proceedings of the IEEE conference on computer vision and pattern recognition, 2018, pp. 8769–8778.

[168] O. VINYALS, C. BLUNDELL, T. LILLICRAP, K. KAVUKCUOGLU, AND D. WIERSTRA, *Matching networks for one shot learning*, in Advances in Neural Information Processing Systems, D. Lee, M. Sugiyama, U. Luxburg, I. Guyon, and R. Garnett, eds., vol. 29, Curran Associates, Inc., 2016.

[169] R. VUORIO, S.-H. SUN, H. HU, AND J. J. LIM, *Multimodal model-agnostic meta-learning via task-aware modulation*, 2019.

[170] D. WANG, E. SHELHAMER, S. LIU, B. OLSHAUSEN, AND T. DARRELL, *Tent: Fully test-time adaptation by entropy minimization*, arXiv preprint arXiv:2006.10726, (2020).

[171] G. WANG AND C. DING, *Effortless active labeling for long-term test-time adaptation*, in Proceedings of the Computer Vision and Pattern Recognition Conference, 2025, pp. 25633–25642.

[172] H. WANG AND Z.-H. DENG, *Cross-domain few-shot classification via adversarial task augmentation*, 2021.

[173] J. WANG, C. LAN, C. LIU, Y. OUYANG, T. QIN, W. LU, Y. CHEN, W. ZENG, AND P. S. YU, *Generalizing to unseen domains: A survey on domain generalization*, IEEE transactions on knowledge and data engineering, 35 (2022), pp. 8052–8072.

[174] M. WANG AND W. DENG, *Deep visual domain adaptation: A survey*, Neurocomputing, 312 (2018), pp. 135–153.

[175] Q. WANG, O. FINK, L. VAN GOOL, AND D. DAI, *Continual test-time domain adaptation*, in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2022, pp. 7201–7211.

[176] R. WANG, H. ZUO, Z. FANG, AND J. LU, *Multiple teacher model for continual test-time domain adaptation*, in Australasian Joint Conference on Artificial Intelligence, Springer, 2023, pp. 304–314.

[177] R. WANG, H. ZUO, J. LU, AND Z. FANG, *Towards robustness prompt tuning with fully test-time adaptation for clip's zero-shot generalization*, in Proceedings of the 32nd ACM International Conference on Multimedia, 2024, pp. 8604–8612.

[178] S. WANG, D. ZHANG, Z. YAN, J. ZHANG, AND R. LI, *Feature alignment and uniformity for test time adaptation*, in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2023, pp. 20050–20060.

[179] X. WANG, Y. PENG, L. LU, Z. LU, M. BAGHERI, AND R. M. SUMMERS, *Chestx-ray8: Hospital-scale chest x-ray database and benchmarks on weakly-supervised classification and localization of common thorax diseases*, in Proceedings of the IEEE conference on computer vision and pattern recognition, 2017, pp. 2097–2106.

[180] Y. WANG, J. HONG, A. CHERAGHIAN, S. RAHMAN, D. AHMEDT-ARISTIZABAL, L. PETERSSON, AND M. HARANDI, *Continual test-time domain adaptation via dynamic sample selection*, in Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, 2024, pp. 1701–1710.

[181] Y. WANG, Q. YAO, J. T. KWOK, AND L. M. NI, *Generalizing from a few examples: A survey on few-shot learning*, ACM computing surveys (csur), 53 (2020), pp. 1–34.

[182] Z. WANG, Y. LUO, L. ZHENG, Z. CHEN, S. WANG, AND Z. HUANG, *In search of lost online test-time adaptation: A survey*, International Journal of Computer Vision, (2024), pp. 1–34.

[183] Z. WANG, E. YANG, L. SHEN, AND H. HUANG, *A comprehensive survey of forgetting in deep learning beyond continual learning*, IEEE Transactions on Pattern Analysis and Machine Intelligence, (2024).

[184] J. WESTON, F. RATLE, AND R. COLLOBERT, *Deep learning via semi-supervised embedding*, in Proceedings of the 25th international conference on Machine learning, 2008, pp. 1168–1175.

[185] G. WILSON AND D. J. COOK, *A survey of unsupervised deep domain adaptation*, ACM Transactions on Intelligent Systems and Technology (TIST), 11 (2020), pp. 1–46.

[186] Q. WU, X. YUE, AND A. SANGIOVANNI-VINCENTELLI, *Domain-agnostic test-time adaptation by prototypical training with auxiliary data*, in NeurIPS 2021 Workshop on Distribution Shifts: Connecting Methods and Applications, 2021.

[187] Y. WU AND K. HE, *Group normalization*, in Proceedings of the European conference on computer vision (ECCV), 2018, pp. 3–19.

[188] Q. XIE, M.-T. LUONG, E. HOVY, AND Q. V. LE, *Self-training with noisy student improves imagenet classification*, in Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, 2020, pp. 10687–10698.

[189] S. XIE, R. GIRSHICK, P. DOLLÁR, Z. TU, AND K. HE, *Aggregated residual transformations for deep neural networks*, in Proceedings of the IEEE conference on computer vision and pattern recognition, 2017, pp. 1492–1500.

[190] X. XIE, J. NIU, X. LIU, Z. CHEN, S. TANG, AND S. YU, *A survey on incorporating domain knowledge into deep learning for medical image analysis*, Medical Image Analysis, 69 (2021), p. 101985.

[191] Y. XIN, J. YANG, S. LUO, H. ZHOU, J. DU, X. LIU, Y. FAN, Q. LI, AND Y. DU, *Parameter-efficient fine-tuning for pre-trained vision models: A survey*, arXiv preprint arXiv:2402.02242, (2024).

[192] H. XU, L. LIU, S. ZHI, S. FU, Z. SU, M.-M. CHENG, AND Y. LIU, *Enhancing information maximization with distance-aware contrastive learning for source-free cross-domain few-shot learning*, IEEE Transactions on Image Processing, 33 (2024), pp. 2058–2073.

[193] H. XU, S. ZHI, AND L. LIU, *Cross-domain few-shot classification via inter-source stylization*, in 2023 IEEE International Conference on Image Processing (ICIP), IEEE, 2023, pp. 565–569.

[194] H. XU, S. ZHI, S. SUN, V. PATEL, AND L. LIU, *Deep learning for cross-domain few-shot visual recognition: A survey*, ACM Computing Surveys, 57 (2025), pp. 1–37.

[195] Y. XU, L. WANG, Y. WANG, C. QIN, Y. ZHANG, AND Y. FU, *Memrein: Rein the domain shift for cross-domain few-shot learning*, (2022).

[196] J. YANG, X. ZHU, A. BULAT, B. MARTINEZ, AND G. TZIMIROPOULOS, *Knowledge distillation meets open-set semi-supervised learning*, International Journal of Computer Vision, 133 (2025), pp. 315–334.

[197] X. YANG, X. CHEN, M. LI, K. WEI, AND C. DENG, *A versatile framework for continual test-time domain adaptation: Balancing discriminability and generalizability*, in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2024, pp. 23731–23740.

[198] X. YANG, Y. GU, K. WEI, AND C. DENG, *Exploring safety supervision for continual test-time domain adaptation*, in Proceedings of the Thirty-Second International Joint Conference on Artificial Intelligence, 2023, pp. 1649–1657.

[199] M. YAZDANPANAH AND P. MORADI, *Visual domain bridge: A source-free domain adaptation for cross-domain few-shot learning*, in Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, 2022, pp. 2868–2877.

[200] Z. YE, J. WANG, H. LIU, Y. ZHANG, AND W. LI, *Adaptive domain-adversarial few-shot learning for cross-domain hyperspectral image classification*, IEEE Transactions on Geoscience and Remote Sensing, 61 (2023), pp. 1–17.

[201] F. YOU, J. LI, AND Z. ZHAO, *Test-time batch statistics calibration for covariate shift*, arXiv preprint arXiv:2110.04065, (2021).

[202] L. YUAN, B. XIE, AND S. LI, *Robust test-time adaptation in dynamic scenarios*, in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2023, pp. 15922–15932.

[203] W. YUAN, T. MA, H. SONG, Y. XIE, Z. ZHANG, AND L. MA, *Both comparison and induction are indispensable for cross-domain few-shot learning*, in 2021 IEEE International Conference on Multimedia and Expo (ICME), IEEE Computer Society, 2021, pp. 1–6.

[204] S. ZAGORUYKO AND N. KOMODAKIS, *Wide residual networks*, arXiv preprint arXiv:1605.07146, (2016).

[205] K. ZAMAN, M. SAH, C. DIREKOGLU, AND M. UNOKI, *A survey of audio classification using deep learning*, IEEE access, 11 (2023), pp. 106620–106649.

[206] L. ZENG, J. HAN, L. DU, AND W. DING, *Rethinking precision of pseudo label: Test-time adaptation via complementary learning*, Pattern Recognition Letters, 177 (2024), pp. 96–102.

[207] X. ZHAI, X. WANG, B. MUSTAFA, A. STEINER, D. KEYSERS, A. KOLESNIKOV, AND L. BEYER, *Lit: Zero-shot transfer with locked-image text tuning*, CVPR, (2022).

[208] D. ZHANG, D. GATICA-PEREZ, S. BENGIO, AND I. MCCOWAN, *Semi-supervised adapted hmms for unusual event detection*, in 2005 IEEE computer society conference on computer vision and pattern recognition (CVPR'05), vol. 1, IEEE, 2005, pp. 611–618.

[209] J. ZHANG, L. QI, Y. SHI, AND Y. GAO, *Domainadaptor: A novel approach to test-time adaptation*, in Proceedings of the IEEE/CVF International Conference on Computer Vision, 2023, pp. 18971–18981.

[210] J. ZHANG, Y. WANG, X. YANG, AND E. ZHU, *A fully test-time training framework for semi-supervised node classification on out-of-distribution graphs*, ACM Transactions on Knowledge Discovery from Data, 18 (2024), pp. 1–19.

[211] M. ZHANG, S. LEVINE, AND C. FINN, *Memo: Test time robustness via adaptation and augmentation*, Advances in Neural Information Processing Systems, 35 (2022), pp. 38629–38642.

[212] Q. ZHANG, Y. JIANG, AND Z. WEN, *Tacdfsl: Task adaptive cross domain few-shot learning*, Symmetry, 14 (2022), p. 1097.

[213] T. ZHANG, Q. CAI, F. GAO, L. QI, AND J. DONG, *Exploring cross-domain few-shot classification via frequency-aware prompting*, arXiv preprint arXiv:2406.16422, (2024).

[214] A. ZHAO, M. DING, Z. LU, T. XIANG, Y. NIU, J. GUAN, AND J.-R. WEN, *Domain-adaptive few-shot learning*, in Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), January 2021, pp. 1390–1399.

[215] B. ZHAO, C. CHEN, AND S.-T. XIA, *Delta: degradation-free fully test-time adaptation*, arXiv preprint arXiv:2301.13018, (2023).

[216] S. ZHAO, B. LI, P. XU, AND K. KEUTZER, *Multi-source domain adaptation in the deep learning era: A systematic survey*, arXiv preprint arXiv:2002.12169, (2020).

[217] B. ZHOU, A. LAPEDRIZA, A. KHOSLA, A. OLIVA, AND A. TORRALBA, *Places: A 10 million image database for scene recognition*, IEEE transactions on pattern analysis and machine intelligence, 40 (2017), pp. 1452–1464.

[218] F. ZHOU, P. WANG, L. ZHANG, Z. CHEN, W. WEI, C. DING, G. LIN, AND Y. ZHANG, *Meta-exploiting frequency prior for cross-domain few-shot learning*, in arXiv preprint arXiv:2411.01432, 2024.

[219] K. ZHOU, Z. LIU, Y. QIAO, T. XIANG, AND C. C. LOY, *Domain generalization: A survey*, IEEE transactions on pattern analysis and machine intelligence, 45 (2022), pp. 4396–4415.

[220] K. ZHOU, J. YANG, C. C. LOY, AND Z. LIU, *Learning to prompt for vision-language models*, International Journal of Computer Vision, 130 (2022), pp. 2337–2348.

[221] J. ZHUANG, B. GONG, L. YUAN, Y. CUI, H. ADAM, N. DVORNEK, S. TATIKONDA, J. DUNCAN, AND T. LIU, *Surrogate gap minimization improves sharpness-aware training*, ICLR, (2022).