Trustworthy Machine Learning under Distribution Shifts

ZHUO HUANG

Supervisor: Assoc. Prof. Tongliang Liu
Auxiliary Supervisor: Asst. Prof. Baosheng Yu

A thesis submitted in fulfilment of
the requirements for the degree of
Doctor of Philosophy

School of Computer Science
Faculty of Engineering
The University of Sydney
Australia

3 March 2026

Abstract

Machine Learning (ML) has been a foundational topic in artificial intelligence (AI), providing both theoretical groundwork and practical tools for its exciting advancements. From ResNet for visual recognition to Transformer for vision-language alignment, the AI models have achieved superior capability to humans. Furthermore, the scaling law has enabled AI to initially develop general intelligence, as demonstrated by Large Language Models (LLMs). To this stage, AI has had an enormous influence on society and yet still keeps shaping the future for humanity.

However, distribution shift remains a persistent "Achilles' heel", fundamentally limiting the reliability and general usefulness of ML systems. As AI becomes increasingly integrated into real-world decision-making and societal infrastructures, the complexity of the problems we ask it to solve continues to grow. These complex environments naturally introduce diverse and unpredictable distribution shifts, which can severely degrade model performance.

Moreover, generalization under distribution shift would also cause trust issues for AIs. For instance, when employing medical AIs across regions, they might perform unsatisfactorily and cause harm. Thus, we also consider the responsibility of AI, i.e., the Trustworthiness of ML, aiming to enhance reliability rather than merely focusing on accuracy.

Motivated by these challenges, my research focuses on *Trustworthy Machine Learning under Distribution Shifts*, with the goal of expanding AI's robustness, versatility, as well as its responsibility and reliability. We carefully study the three common distribution shifts into: (1) Perturbation Shift, (2) Domain Shift, and (3) Modality Shift. For all scenarios, we also rigorously investigate trustworthiness via three aspects: (1) Robustness, (2) Explainability, and (3) Adaptability. Based on these dimensions, we propose effective solutions and fundamental insights, meanwhile aiming to enhance the critical ML problems, such as efficiency, adaptability, and safety.

Statement of Originality

I certify that the work in this thesis has not previously been submitted for a degree, nor has it been submitted as part of the requirements for a degree except as fully acknowledged within the text.

I also certify that the thesis has been written by me. Any help that I have received in my research work and the preparation of the thesis itself has been acknowledged. In addition, I certify that all information sources and literature used are indicated in the thesis.

Zhuo Huang
School of Computer Science
Faculty of Engineering
The University of Sydney

30 Dec 2025

Statement of Funding

I acknowledge my PhD is supported by The JD Technology Scholarship for Postgraduate Research in Artificial intelligence.

Zhuo Huang
School of Computer Science
Faculty of Engineering
The University of Sydney 30 Dec 2025

Statement of Generative AI

During the preparation of this thesis, ChatGPT (OpenAI) was used for the purposes of text enhancement. The use of this generative AI tool includes minor paraphrasing, sentence restructuring, and spelling correction in several draft chapters. I confirm that where text was modified by generative AI, the content was reviewed for possible errors, inaccuracies, and bias. The author takes full responsibility for the submitted thesis and ensures the work is their own and has used generative AI within the intended purpose of use.

Zhuo Huang
School of Computer Science
Faculty of Engineering
The University of Sydney

30 Dec 2025

Authorship Attribution Statement

This thesis was conducted at The University of Sydney, under the supervision of Assoc. Prof. Tongliang Liu, between 2022 and 2025. The main results presented in this dissertation were first introduced in the following publications:

(1) **Zhuo Huang**, Xiaobo Xia, Li Shen, Bo Han, Mingming Gong, Chen Gong, Tongliang Liu. "Harnessing Out-Of-Distribution Examples via Augmenting Content and Style". In *The Eleventh International Conference on Learning Representations (ICLR)*, 2023. Presented in Chapter 2. I identified the research problem, formulated setting, implement the experiments, and drafted the manuscript. The polish of the paper is helped by other co-authors.

(2) **Zhuo Huang**, Miaoxi Zhu, Xiaobo Xia, Li Shen, Jun Yu, Chen Gong, Bo Han, Bo Du, Tonglinag Liu "Robust Generalization against Photon-Limited Corruptions via Worst-Case Sharpness Minimization". *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023. Presented in Chapter 3. I identified the research problem, formulated setting, implement the experiments, and drafted the manuscript. The theoretical formulation and polish of the paper are helped by other co-authors.

(3) **Zhuo Huang**, Muyang Li, Li Shen, Jun Yu, Chen Gong, Bo Han, Tongliang Liu. "Winning Prize Comes from Losing Tickets: Improve Invariant Learning by Exploring Variant Parameters for Out-of-Distribution Generalization". In *International Journal of Computer Vision (IJCV)*, 2024. Presented in Chapter 4. I identified the research problem, formulated setting, implement the experiments, and drafted the manuscript. The polish of the paper is helped by other co-authors.

(4) **Zhuo Huang**, Chang Liu, Yinpeng Dong, Hang Su, Shibao Zheng, Tongliang Liu. "Machine Vision Therapy: Multimodal Large Language Models Can Enhance Visual Robustness via Denoising In-Context Learning". In *International Conference on Machine Learning (ICML)*, 2024. Presented in Chapter 5. I identified the

research problem, formulated setting, implement the experiments, and drafted the manuscript. The polish of the paper and some experiments are helped by other co-authors.

(5) **Zhuo Huang**, Gang Niu, Bo Han, Masashi Sugiyama, Tongliang Liu. "Towards Out-of-Modal Generalization without Instance-level Modal Correspondence". In *The Thirteenth International Conference on Learning Representations (ICLR)*, 2025. Presented in Chapter 6. I identified the research problem, formulated setting, implement the experiments, and drafted the manuscript. The polish of the paper and problem setting are helped by other co-authors.

In addition to the statements above, in cases where I am not the corresponding author of a published item, permission to include the published material has been granted by the corresponding author.

Zhuo Huang
School of Computer Science
Faculty of Engineering
The University of Sydney 30 Dec 2025

As supervisor for the candidature upon which this thesis is based, I can confirm that the authorship attribution statements above are correct.

Tongliang Liu
School of Computer Science
Faculty of Engineering
The University of Sydney 30 Dec 2025

Sidere mens eadem mutato.

Acknowledgements

During my PhD study, there are many people has been a great help to me.

First of all, I would like to thank my PhD supervisor, Prof. Tongliang Liu, who has been very supportive in every part of this thesis. I feel very grateful for his guidance and consistent patience. I can focus on my research and freely discover what interests me has been a great blessing, largely owes to him. I will always remember his wisdom and insights on critical thinking and even life philosophy. His calm personality has largely shaped my attitude towards ups and downs in PhD and life. I sincerely express my appreciation to thank him for everything.

I also wish to express my deep appreciation to Prof. Masashi Sugiyama and Dr. Gang Niu, who have been very supportive when I was visiting RIKEN AIP in Tokyo, and I enjoy every inspiring conversion we have. I also want to thank Prof. Chen Gong, A/Prof. Mingming Gong, A/Prof. Bo Han, and A/Prof. Li Shen, who have been my encouraging mentors and collaborators. I also wish to thank A/Prof. Hang Su, Dr. Yinpeng Dong, Dr. Chang Liu, and Dr. Zijian Zhu for helping me when I was visiting Tsinghua University. I also want to thank some of the brilliant minds that I have learned from: Prof. Kun Zhang, Prof. Javen Qinfeng Shi, Prof. Lina Yao, Dr. Dadong Wang, A/Prof. Shuo Chen, A/Prof. Takashi Ishida, Dr. Feng Liu, Dr. Dong Gong, and Dr. Zhen Fang. Their encouragement and advice have been a priceless treasure during my research life, I wish to express my sincere appreciation to them all.

Moreover, I would like to thank my precious collaborators at TML group: Dr. Yu Yao, Dr. Xiaobo Xia, Dr. Songhua Wu, Dr. Huaxi Huang, Dr. Yingbin Bai, Dr. Dawei Zhou, Dr. Chaojian Yu, Dr. Runnan Chen, Dr. Liangcheng Liu, Dr. Yuhao Wu, Dr. Cong Lei, Dr. Runqi Lin, Dr. Vincent Qu, Jialiang Shen, Zhaoqing Wang, Suqin Yuan, Jiyang Zheng, Yexiong Lin, Muyang Li, Xiuchuan Li, Ziming Hong, Yongli Xiang, Li He, Jiaxin Huang, Tianyu Huang, Xiangyu Sun, Quan Tran, Zhen Huang, Andrew Cao, Longjie Zhao, Peng

Mi, Zhenchen Wan, Jun Wang, Keshen Zhou, and Yuxiang Zhen. They are extraordinary people who made my PhD life unforgettable. I show my gratitude to them from the bottom of my heart.

Further, I want thank my dear friends: Dr. Qizhou Wang, Dr. Erdun Gao, Dr. Weijian Deng, Weijie Tu, Chengyi Cai, Dr. Jianing Zhu, Puning Yang, Xiaohao Liu, Jiacheng Zhang, Zhongyi Bai, Xiaotong Yu, Dr. Hongyu Zhou, Tinghui Li, Dr. Jiakun Yu, Dr. Andong Wang, Dr. Zhen-Yu Zhang, Prof. Ximing Li, Dr. Ming-Kun Xie, Dr. Xin-Qiang Cai, Dr. Wei Wang, Dr. Yuning Qiu, Dr. Haonan Huang, Dr. Mingyuan Bai, Miaoxi Zhu, Hanlue Zhang, Zhengqing Gao, Ziwen Li, Dr. Lin Li, Haoyu Wang, and Dr. Jihyo Kim. Their kindness and care have accompanied me through this journey, and I could not express my appreciation enough to them.

Last but not least, I want to express my deepest gratitude to my parents, Yucheng Huang and Hongying Yang. They are the most important people in my life, who are constantly there for me and encourage me to step forward. No words can show my love and gratitude for them.

I dedicate this thesis to them.

Contents

List of Figures

List of Tables

CHAPTER 1

Introduction

1.1 Background

Machine learning (ML) is one of the most effective approaches to Artificial Intelligence (AI), which leverages statistical tools to find, recognize, and utilize patterns in data, it has provided theoretical ground and practical tools that have led to the flourishing of AI. As a result, it has achieved tremendous success in a wide spectrum of applications, such as computer vision, natural language processing, speech recognition, autonomous driving, and robotics, significantly influencing human civilization.

The early AI revolution is led by deep neural networks [1], such as AlexNet [2], ResNet [3], and DenseNet [4], which have demonstrated that AI can achieve superior performance to human intelligence in visual recognition tasks. In 2017, Google introduced the Transformer architecture [5], which enabled models to handle vast amounts of text [6] as well as visual tokens [7] in parallel, thereby scaling up the magnitude of information processing. Based on such an architecture, it has become a common understanding that simply expanding the learning scale can always achieve improved performance, also known as the *Scaling Law* [8]. As a result, Large Language Models (LLMs) such as GPT [9], Llama [10], and Qwen [11] have been scaled-up to Billions of parameters, along with many emergent capabilities including In-Context Learning, Chain-of-Thought Reasoning, and Arithmetic Operation. Beyond language-only models, Multi-modal LLMs (MLLMs) that combine language and vision to allow advanced human-computer interaction have significantly increased their real-world capabilities, such as ChatGPT [12], Gemini [13], Claude [14], and DeepSeek [15]. Until now, most commercialized AI models are MLLMs and they play a vital role in boosting

efficiency and productivity of daily tasks, such as document organization, software development, content creation and design.

Despite its success and real-world impact, existing learning models still heavily rely on the independent and identically distributed (IID) assumption. Particularly, ML aims to learn from training data, which is collected to simulate the expected real-world environment. Based on the Empirical Risk Minimization (ERM) framework [16, 17], ML is guaranteed to generalize when tested under a similar data distribution to the training distribution. Therefore, traditional ML models such as Support Vector Machine (SVM) [18] and Multi-Layer Perceptron (MLP) [19] can work satisfactorily under simple scenarios such as fraud detection and spam filtering.

However, due to domain shift, dataset bias, or evolving environments in real-world scenarios, the IID assumption is constantly violated, introducing Out-of-Distribution (OOD) data that hinders the effectiveness of ML. For example, autonomous vehicles are trained on thousands of hours of videos from sunny, dry, and daytime conditions, when it is deployed at rainy, wet, and nighttime streets, they might have serious performance malfunctioning; a model trained in 2022 is asked who is Prime Minister of UK, but it won't give the correct answer because the answer is changed over time. Therefore, when the distribution of test dataset is shifted from the training dataset, it introduces significant complexity and variance to the learning process. Thus, the effectiveness of many existing ML methods is suboptimal in practice, which raises serious concerns on their capability under realistic distribution shift.

Apart from the generalization capability of AI, it is also crucial to enhance its responsibility. To achieve this, Trustworthy Machine Learning (TML) [20] was proposed to focus on reliability and integrity of AI systems rather than focusing primarily on accuracy, which has attracted abundant attention from both academy and industry. For example, OpenAI pioneered AI Alignment for Human and proposed Reinforcement Learning from Human Feedback (RLHF) [21] which aims to ensure the value and morality of AIs are well aligned with human; moreover, institutions like UC Berkeley, NYU, and Mila launched MATS [22] program aiming to prevent the possibility of models accidentally learning dangerous behaviors in tasks, such as writing code and generating images.

In the era of LLMs and the dawn of general intelligence, it is urgent to consider how we can live with AI and how AI can be aligned with us. Otherwise, as it gets embodied into human society, the consequences of untrustworthy AIs would be devastating, causing irreparable economic losses and social disaster. In particular, AIs that provide customer services might be "jailbroken" by malicious users and be forced to make improper deals, leading to massive monetary cost; moreover, when LLMs scrape numerous web data for pre-training, it might accidentally collect user-sensitive data and violate the privacy protocol, posing severe safety and security concerns.

However, given the significance of the above two fundamental aspects, there lack of a systematic study to address them simultaneously. In fact, the capability for AI under distribution shifts and its responsibility for trustworthiness can conflict with each other. Intuitively, enhancing generalization beyong IID data enforces AI to greedily obsorbing unknown knowledge and autonomously handle uncertainty, e.g., for treatment works for over 50 men in a trial, over-generalized AI might assert that this treatment is effective. As a result, such an excessive agency would eventually reach domains that are misaligned with human values, thus damaging the reliability of AI. On the other hand, heavy penalization to achieve strict trustworthiness would make AI overcautious, which could fail to perform under challenging conditions, e.g., a self-driving car is trained for millions of miles with almost zero risk, it could just stop and stay stationary when it is deployed on snowy mountain paths because it detects danger in making any action. Therefore, trading off between capability and responsibility will continue to be one critical challenge in the future of AI.

In this thesis, motivated by the goal of developing general intelligence for the best human benefits, we carefully investigate trustworthy machine learning under realistic distribution shift problems, aiming to find solutions to ensure capability and responsibility simultaneously. To achieve this, we explore three perspectives for both distribution shift and trustworthiness, as shown in Figure 1.1. Particularly, to investigate responsibility, we consider:

- Robustness, which demonstrates the ability of AI to maintain stable and effective performance under challenging scenarios.

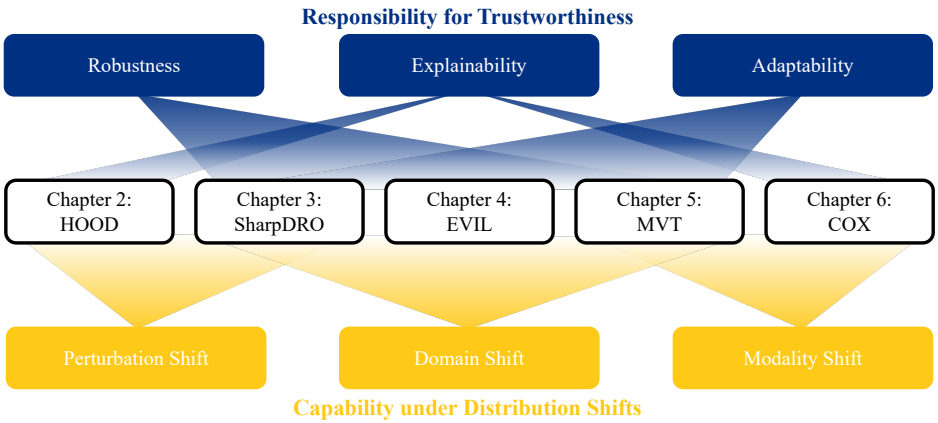


FIGURE 1.1. Outline for trustworthy machine learning under distribution shifts. We explore Robustness, Explainability, and Adaptability for trustworthiness of ML models, and consider three types of common distribution shifts, including Perturbation Shift, Domain Shift, and Modality Shift. From Chapter 2 to Chapter 6, we shed light on different trustworthy aspects and distribution shift scenarios by solving realistic tasks and proposing effective approaches with theoretical grounding.

- Explainability, which aims to understand the nature of the distribution shift and make the decision-making process interpretable.
- Adaptability, which denotes the ability of AI to perform autonomously across various conditions without human intervention.

These three aspects of TML are highly vulnerable or essential for studying generalization under distribution shifts; thus, we focus on them for this thesis. Note that there are other topics such as fairness and privacy in TML, but they align orthogonal from generalization under distribution shift. The success can be achieved by combining their findings, such as reweighting [23] or unlearning [24]. Moreover, to evaluate the generalization capability under distribution shifts, we consider:

- Perturbation Shift, where natural or synthetic noise is applied to the dataset, largely deviating its statistical patterns.
- Domain Shift, where data are collected from different environments, further introducing confounding factors that significantly mislead the prediction.
- Modality Shift, where the knowledge is represented in a heterogeneous structure, intensifying the feature incompatibility and hindering the learning feasibility.

The above distribution shifts demonstrate different intensities of shifts. Perturbation shift is the mild shift, as it has the same data type from the same environment; domain shift is moderate because it is caused by different environmental factors; and modality shift is the most extreme one because it has a different data type, implying a change of input space. Yet, they are ubiquitous and all belong to the Covariate Shift [25] in statistical ML. There are other less common shifts, including Target Shift [26] and Concept Shift [27]. Specifically, target shift is caused by the change of label distribution, but given the simpicity of the label space in most applications, it can be easily solved via reweighting or resampling; concept shift is less-studied due to that concepts and knowledge in real-world applications are generally stationary. Thus, they are excluded from the scope of this thesis. Next, we will summarize the contributions of this thesis and demonstrate the structure by chapter.

1.2 Summary of Contributions

To sum up, this thesis focuses on Trustworthy Machine Learning under Distribution Shift, aiming to explore responsibility and capability simultaneously under practical scenarios. By considering various levels of shift, my studies cover a wide spectrum of applications and provide insights for traditional topics as well as futuristic directions. Additionally, we provide extensive analyses and theoretical groundings to rigorously justify my research under extensive datasets, models, and settings, validating their trustworthiness along with many other notable benefits, such as efficiency, safety, and affordability. Below, we reveal the logical structure of the following chapters and emphasize their contributions.

In Chapter 2, we observe that OOD data present a double-edged effect that could either enhance or harm generalization in different scenarios. Therefore, we aim to understand such an effect and harness OOD data effectively for various scenarios. We inspect from a causal inference perspective to explain the data generation process, which reveals that image data can be decomposed into “content” and “style”. Beneficial OOD data is caused by a style change, and harmful OOD data is caused by a content change. We propose Understanding and Harnessing OOD data (HOOD) framework that disentangles content and style during inference,

then separately creates beneficial and harmful OOD data using learnable adversarial perturbations. By training on such tailor-designed examples, we can enhance the generalization performance and properly handle OOD data in the wild.

Distribution shifts are not only presented in different types, but they also appear in various strengths, forming a traceable distribution. In Chapter 3, we study learning under corruptions that follow real-world noise distribution. Intuitively, corruption is caused by multiple noise elements which occur with a certain probability during a time interval, knwon as "photon-limited corruptions". By modeling it via Poisson distirbution, the strengths levels of the corruption can be identified. To deal with such a multi-strength distribution shift, we propose Sharpness-based Distribution Robust Optimization (SharpDRO) to reweight each distirbution and enhance its generalization by encouraging the loss smoothness, i.e., minimizing "sharpness". As a result, my approaches perform effectively and stably across various corruption strengths, demonstrating robustness and adaptability in practical applications.

Further, to deploy a TML model in the real world, it needs to autonomously identify the essential features without latching onto the confounding information. In Chapter 4, we explore neural regeneration where AIs can dynamically adapt across different environments and extract domain invariant knowledge. To solve this problem, we propose Exploring Variant parameters for Invariant Learning (EVIL) framework, which identifies domain invariant features by growing corresponding neural paths. Moreover, we enhance such a process by further incorporating domain variant knowledge simultaneously during training, which in turn helps identify the variant parameters. By dropping such undesirable neural paths, models can be less affected by domain shift, thus showing enhanced capability and adaptability. Additionally, the proposed approach is conducted via sparse training, thus it also demonstrates great efficiency.

Existing distribution shift studies are mainly focused on perturbations and domain shifts. However, knowledge adaptation across different modalities is rarely explored. Therefore, in Chapter 5, we make an initial attempt at generalization across the two most common modalities, i.e., language and vision. Learning from language has been very successful thanks to

the development of LLMs, which already demonstrate general-purpose capabilities. For vision tasks, it is highly dependent on human supervision, which is very resource-consuming. Therefore, a question is asked: Is it possible to replace human supervision by AI supervision? Hence, we propose Machine Vision Therapy (MVT) framework that leverages LLMs to guide the visual learning process. Due to the modality gap between language models and vision models, we propose to conduct In-Context Learning (ICL) to analyze visual tokens in textual contexts. As a result, the proposed MVT can largely boost the visual robustness, successfully adapting knowledge across modalities.

In Chapter 6, instead of focusing on generalization across common modalities, we propose a novel problem named Out-of-Modal (OOM) Generalization, which aims to leverage well-known modal knowledge to infer rare modalities. For example, language and vision modalities are quite extensive, but rare modalities, such as tactile, LiDAR, and genomics, are not well-explored by AIs. Due to the difficulty and cost of collecting and organizing such data, it is extremely challenging to conduct large-scale training. Therefore, exploiting the hidden knowledge from rare modalities based on the existing ones is essential for general intelligence. In OOM generalization, we first understand the modality information via the perspective of multimodal interaction, providing insights for extracting generalizable knowledge. By applying to real-world settings, it is possible to comprehend a novel modality from scratch, paving the future directions for general-purpose AI.

Additionally, we provide theoretical proofs for all the theoretical studies in this thesis in Chapter 7. At last, we conclude all our contributions and provide prospective discussions for future studies in Chapter 8.

CHAPTER 2

Understanding and Harnessing OOD Data

Machine learning models are vulnerable to Out-Of-Distribution (OOD) examples, and such a problem has drawn much attention. However, current methods lack a full understanding of different types of OOD data: there are *benign* OOD data that can be properly adapted to enhance the learning performance, while other *malign* OOD data would severely degenerate the classification result. To understand and Harness OOD data, this Chapter proposes a HOOD method that can leverage the *content* and *style* from each image instance to identify benign and malign OOD data. Particularly, we design a variational inference framework to causally disentangle content and style features by constructing a structural causal model. Subsequently, we augment the content and style through an intervention process to produce malign and benign OOD data, respectively. The benign OOD data contain novel styles but hold our interested contents, and they can be leveraged to help train a style-invariant model. In contrast, the malign OOD data inherit unknown contents but carry familiar styles, by detecting them can improve model robustness against deceiving anomalies. Thanks to the proposed novel disentanglement and data augmentation techniques, HOOD can effectively deal with OOD examples in unknown and open environments, whose effectiveness is empirically validated in three typical OOD applications including OOD detection, open-set semi-supervised learning, and open-set domain adaptation.

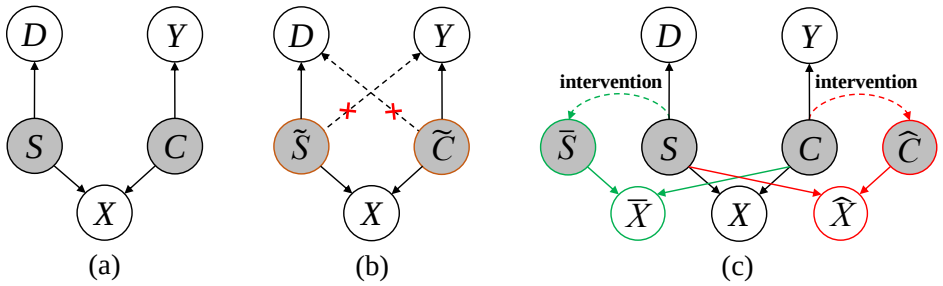


FIGURE 2.1. (a) An ideal causal diagram denoting the data generating process. (b) Illustration of our disentanglement. The brown-edged variables $\tilde{C}$ and $\tilde{S}$ are approximations of content C and style S. The dashed lines indicate the unwanted causal relations to be broken. (c) Illustration of the data augmentation of HOOD. The green lines and red lines denote the augmentation of benign OOD data $\bar{X}$ and malign OOD data $\hat{X}$, respectively. In all figures, the blank variables are observable and the shaded variables are latent.

2.1 Introduction

Learning in the presence of Out-Of-Distribution (OOD) data has been a challenging task in machine learning, as the deployed classifier tends to fail if the unseen data drawn from unknown distributions are not properly handled [28, 29]. Such a critical problem ubiquitously exists when deep models meet domain shift [30, 31] and unseen-class data [28, 32], which has drawn a lot of attention in some important fields such as OOD detection [33, 28, 34, 35, 36, 37, 38, 39], Open-Set Domain Adaptation (DA) [40, 41], and Open-Set Semi-Supervised Learning (SSL) [42, 43, 44, 45, 46, 47, 48, 49].

In the above fields, OOD data can be divided into two types, namely *benign* OOD data[1] and *malign* OOD data. The benign OOD data can boost the learning performance on the target distribution through DA techniques [51, 31], but they can be misleading if not being properly exploited. To improve model generalization, many *positive data augmentation* techniques [52, 53] have been proposed. For instance, the performance of SSL [54, 55] has been greatly improved thanks to the augmented benign OOD data. On the contrary, malign OOD data with unknown classes can damage the classification results, but they are deceiving and hard to detect [28, 35, 56, 57]. To train a robust model against malign OOD data, some works [58, 59] conduct *negative data augmentation* to generate "hard" malign data

[1]We follow [50] to regard the augmented data as a type of OOD data

which resemble in-distribution (ID) data. By separating such "hard" data from ID data, the OOD detection performance can be improved. When presented with both malign and benign OOD data, it is more challenging to decide which to separate and which to exploit. As a consequence, the performance of existing open-set methods could be sub-optimal due to two drawbacks: (1) radically exploiting too much malign OOD data, and (2) conservatively denying too much benign OOD data.

In this Chapter, we propose a HOOD framework (see Figure 2.2) to properly harness OOD data in several OOD problems. To distinguish benign and malign OOD data, we model the data generating process by following the structural causal model (SCM) [60, 61, 62] in Figure 2.1 (a). Particularly, we decompose an image instance X into two latent components: (1) *content* variable C which denotes the interested object, and (2) *style* variable S which contains other influential factors such as brightness, orientation, and color. The content C can indicate its true *class* Y, and the style S is decisive for the environmental condition, which is termed as *domain* D. Intuitively, malign OOD data cannot be incorporated into network training, because they contain unseen contents, thus their true classes are different from any known class; and benign OOD data can be adapted because they only have novel styles but contain the same contents as ID data. Therefore, we can distinguish the benign and malign OOD data based on the extracted the content and style features.

In addition, we conduct causal disentanglement through maximizing an approximated evidence lower-bound (ELBO) [63, 64, 65] of joint distribution $P(X, Y, D)$. As a result, we can effectively break the spurious correlation [61, 60, 66, 67, 68] between content and style which commonly occurs during network training [69], as shown by the dashed lines in Figure 2.1 (b). In the ablation study, we find that HOOD can correctly disentangle content and style, which can correspondingly benefit generalization tasks (such as open-set DA and open-set SSL) and detection task (such as OOD detection).

To further improve the learning performance, we conduct both positive and negative data augmentation by solely intervening the style and content, respectively, as shown by the blue and red lines in Figure 2.1 (c). Such process is achieved through backpropagating the gradient computed from an intervention objective. As a result, style-changed data $\bar{X}$ must be

identified as benign OOD data, and content-changed data $\hat{X}$ should be recognized as malign OOD data. Without including any bias, the benign OOD data can be easily harnessed to improve model generalization, and the malign OOD data can be directly recognized as harmful ones which benefits the detection of unknown anomalies. By conducting extensive experiments on several OOD applications, including OOD detection, open-set SSL, and open-set DA, we validate the effectiveness of our method on typical benchmark datasets. To sum up, our contributions are three-fold:

- We propose a unified framework dubbed HOOD which can effectively disentangle the content and style features to break the spurious correlation. As a result, benign OOD data and malign OOD data can be correctly identified based on the disentangled features.
- We design a novel data augmentation method which correspondingly augments the content and style features to produce benign and malign OOD data, and further leverage them to enhance the learning performance.
- We experimentally validate the effectiveness of HOOD on various OOD applications, including OOD detection, open-set SSL, and open-set DA.

2.2 Related Work

OOD applications contains three typical problems, namely OOD detection, open-set SSL, and open-set DA. OOD detection [28, 35] aims to train a robust model which can accurately identify the newly-emerged malign OOD data during the test phase. Open-set SSL [70, 71, 72, 73, 49] deals with the problem when labeled data are scarce and the unlabeled data are contaminated by malign OOD data. As for open-set DA [41, 40], it tries to transfer the knowledge from source ID data to the benign OOD data in target domain, meanwhile detecting the malign OOD data that are encountered during transferring. In both three applications, the predictive confidence has been frequently leveraged to separate malign OOD data [28, 35, 74, 75]. Moreover, ID data and OOD data can be distinguished via using a discriminator [58, 76, 49, 77]. Further, various open classifiers are designed to predict OOD dataset as

unknown [78, 79, 41]. Thanks to the advances in unsupervised learning, many approaches employ self-supervised learning to distinguish ID data and OOD data [80, 81, 82].

Causality in OOD problems mainly focuses on learning invariant representations that stay constant when other causal factors are changing, thus achieving better performance when facing non-stationary data distribution. To accomplish this goal, it is common to learn causal factors and non-causal factors through the variational auto-encoder framework [63, 83]. Thanks to which, domain adaptation [84, 85, 26] and domain generalization [86, 87] can be tackled through extracting the domain invariant features. Moreover, based on causal effects, the biased feature can be eliminated through re-weighting [88, 89]. Additionally, the spurious correlation which is harmful for inference could be alleviated through do-calculus [90, 91, 61]. Recent methods [92, 93, 94] conduct data augmentations with self-supervised learning to train a robust model that can handle distribution shifts and corruptions.

In general, HOOD has two major differences from existing methods in OOD applications and causality. On one hand, instead of treating an image instance as a whole as commonly done in many approaches, HOOD can properly leverage OOD examples through their disentangled contents and styles. Moreover, augmenting content and style can help improve generalization and robustness simultaneously. On the other hand, current causal approaches are incapable of dealing with malign OOD data, but HOOD is able to learn style-invariant features from benign OOD data, meanwhile avoiding the damage brought by malign OOD data.

2.3 Methodology

In this section, we propose our HOOD framework as shown in Figure 2.2. Specifically, we utilize the class labels of labeled data and the pseudo labels [95] of unlabeled data as the class supervision to capture the content feature. Moreover, we perform different types of data augmentation and regard the augmentation types as the domain supervision for each style. Thereby, each instance $\mathbf{x}$ is paired with a class label y and a domain label d. Then, we apply two separate encoders g_c and g_s parameterized by θ_c and θ_s to model the posterior distributions $q_{\theta_c}(C \mid X)$ and $q_{\theta_s}(S \mid X)$, respectively. Subsequently, the generated C and S

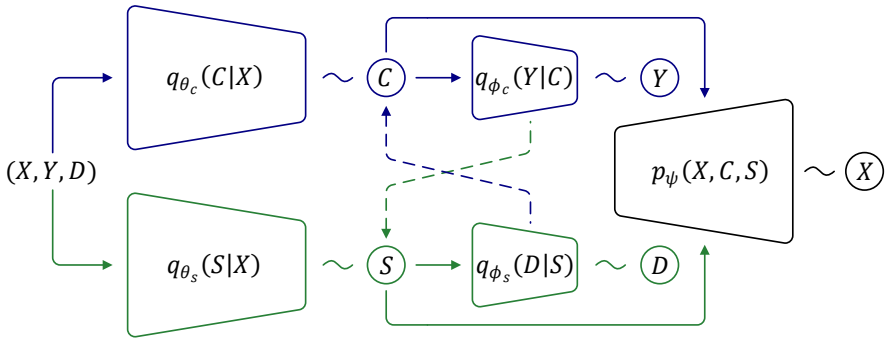


FIGURE 2.2. Architecture of the HOOD. The solid lines denote the inference flow, the dashed lines indicate the disentanglement of content and style, and the tildes stand for the approximation of the corresponding variables.

are correspondingly fed into two fully-connected classifiers f_c and f_s parameterized by ϕ_c and ϕ_s, which would produce the label predictions $q_{\phi_c}(Y \mid C)$ and $q_{\phi_s}(D \mid S)$, respectively. To further enhance the identifiability of C and S, a decoder h with parameter ψ is employed to reconstruct the input instance $\mathbf{x}$ based on its content and style.

Below, we describe the detailed procedures and components during modeling HOOD. We first introduce the proposed variational inference framework for disentangling the content and style based on the constructed SCM. Subsequently, we conduct intervention to produce benign OOD data and malign OOD data. Further, we appropriately leverage the benign and malign OOD data to boost the learning performance. Finally, we formulate the deployment of HOOD in three OOD applications.

2.3.1 Variational Inference for Content and Style Disentanglement

First, we assume that the data generating process can be captured by certain probability distributions. Therefore, according to the constructed SCM in Figure 2.1 (a), the joint distribution $P(X, Y, D, C, S)$ of the interested variables can be factorized as follows:

$$P(X, Y, D, C, S) = P(C, S)P(Y, D \mid C, S)P(X \mid C, S). \tag{2.1}$$

Based on the SCM in Figure 2.1 (a), Y and D are conditionally independent to each other, i.e., $Y \perp\!\!\!\perp D \mid (C, S)$, so we have $P(Y, D \mid C, S) = P(Y \mid C, S)P(D \mid C, S)$. Similarly,

we have $P(C, S) = P(C)P(S)$. Moreover, we can also know that Y is not conditioned on S, and D is not conditioned on C. Hence, we can further derive $P(Y, D \mid C, S) = P(Y \mid C)P(D \mid S)$.

However, the aforementioned spurious correlation frequently appears when facing OOD examples [69]. As a consequence, when variational inference is based on the factorization in Equation (2.1), the approximated content $\tilde{C}$ and style $\tilde{S}$ could both directly influence Y and D, i.e., $Y \leftarrow \tilde{C} \rightarrow D$ and $Y \leftarrow \tilde{S} \rightarrow D$, thus leading to inaccurate approximations. However, the desired condition is $Y \leftarrow C \nrightarrow D$ and $Y \nleftarrow S \rightarrow D$. We can see that the unwanted correlations $\tilde{C} \rightarrow D$ and $\tilde{S} \rightarrow Y$ in Figure 2.1 (b) is caused by erroneous posteriors $P(D \mid \tilde{C})$ and $P(Y \mid \tilde{S})$. Therefore, to break the correlations, the posteriors $q_{\phi_s}(D \mid C)$ and $q_{\phi_c}(Y \mid S)$ which are correspondingly approximated by the decoders ϕ_s and ϕ_c can be used as denominators to $q_{\phi_c}(Y \mid C)$ and $q_{\phi_s}(D \mid S)$, respectively. In this way, we can successfully disentangle content C and style S and ensure the decoding process of Y and D would not be influenced by spurious features from S and C, respectively. To this end, our factorization in Equation (2.1) can be approximated as:

$$\tilde{P}(X, Y, D, C, S) \coloneqq \frac{P(C)P(S)P(Y \mid C)P(D \mid S)P(X \mid C, S)}{q_{\phi_s}(D \mid C)q_{\phi_c}(Y \mid S)}. \tag{2.2}$$

Then, we maximize the log-likelihood of the joint distribution $p(\mathbf{x}, y, d)$ of each data point $(\mathbf{x}, y, d)$:

$$\log p(\mathbf{x}, y, d) \coloneqq \log \int_c \int_s \tilde{p}(\mathbf{x}, y, d, c, s) \mathrm{d}c\mathrm{d}s, \tag{2.3}$$

in which we use lower case to denote the values of corresponding variables. Due to the integration of latents C and S is intractable, we follow variational inference [63] to obtain an

approximated evidence lower-bound $\tilde{ELBO}(\mathbf{x}, y, d)$ of the log-likelihood in Equation (2.3):

$$\begin{aligned}
\log p(\mathbf{x}, y, d) &= \log \int_c \int_s \tilde{p}(\mathbf{x}, y, d, c, s) \mathrm{d}c\mathrm{d}s \\
&= \log \int_c \int_s \tilde{p}(\mathbf{x}, y, d, c, s) \frac{q_\theta(c, s \mid \mathbf{x})}{q_\theta(c, s \mid \mathbf{x})} \mathrm{d}c\mathrm{d}s \\
&= \log \mathbb{E}_{(c,s)\sim q_\theta(C,S|\mathbf{x})} \left[\frac{\tilde{p}(\mathbf{x}, y, d, c, s)}{q_\theta(c, s \mid \mathbf{x})} \right] \\
&\geq \mathbb{E}_{(c,s)\sim q_\theta(C,S|\mathbf{x})} \left[\log \frac{\tilde{p}(\mathbf{x}, y, d, c, s)}{q_\theta(c, s \mid \mathbf{x})} \right] \coloneqq \tilde{ELBO}(\mathbf{x}, y, d).
\end{aligned} \tag{2.4}$$

Recall the modified joint distribution factorization in Equation (2.2), we can have:

$$\begin{aligned}
\tilde{ELBO}(\mathbf{x}, y, d) =& \mathbb{E}_{(c,s)\sim q_\theta(C,S|\mathbf{x})} \left[\log \frac{p(c)p(s)q_{\phi_c}(y \mid c)q_{\phi_s}(d \mid s)p_\psi(\mathbf{x} \mid c, s)}{q_\theta(c, s \mid \mathbf{x})q_{\phi_c}(y \mid s)q_{\phi_s}(d \mid c)} \right] \\
=& \mathbb{E}_{(c,s)\sim q_\theta(C,S|\mathbf{x})} \left[\log \frac{p(c)p(s)}{q_{\theta_c}(c \mid \mathbf{x})q_{\theta_s}(s \mid \mathbf{x})} \right] + \mathbb{E}_{(c,s)\sim q_\theta(C,S|\mathbf{x})} \left[\log \frac{q_{\phi_c}(y \mid c)}{q_{\phi_s}(d \mid c)} \right] \\
&+ \mathbb{E}_{(c,s)\sim q_\theta(C,S|\mathbf{x})} \left[\log \frac{q_{\phi_s}(d \mid s)}{q_{\phi_c}(y \mid s)} \right] + \mathbb{E}_{(c,s)\sim q_\theta(C,S|\mathbf{x})} \left[\log p_\psi(\mathbf{x} \mid c, s) \right] \\
=& \mathbb{E}_{(c)\sim q_{\theta_c}(C|\mathbf{x})} \left[\log \frac{p(c)}{q_{\theta_c}(c \mid \mathbf{x})} \right] + \mathbb{E}_{(s)\sim q_{\theta_s}(S|\mathbf{x})} \left[\log \frac{p(c)p(s)}{q_{\theta_s}(s \mid \mathbf{x})} \right] \\
&+ \mathbb{E}_{(c)\sim q_{\theta_s}(C|\mathbf{x})} \left[\log \frac{q_{\phi_c}(y \mid c)}{q_{\phi_s}(d \mid c)} \right] + \mathbb{E}_{(s)\sim q_{\theta_s}(S|\mathbf{x})} \left[\log \frac{q_{\phi_s}(d \mid s)}{q_{\phi_c}(y \mid s)} \right] \\
&+ \mathbb{E}_{(c,s)\sim q_\theta(C,S|\mathbf{x})} \left[\log p_\psi(\mathbf{x} \mid c, s) \right] \\
=& - KL(q_{\theta_c}(c \mid \mathbf{x}) \| p(C)) - KL(q_{\theta_s}(s \mid \mathbf{x}) \| p(S)) \\
&+ \mathbb{E}_{c\sim q_{\theta_c}(C|\mathbf{x})} \left[\log q_{\phi_c}(y \mid c) - \log q_{\phi_s}(d \mid c) \right] \\
&+ \mathbb{E}_{s\sim q_{\theta_s}(S|\mathbf{x})} \left[\log q_{\phi_s}(d \mid s) - \log q_{\phi_c}(y \mid s) \right] \\
&+ \mathbb{E}_{(c,s)\sim q_\theta(C,S|\mathbf{x})} \left[\log p_\psi(\mathbf{x} \mid c, s) \right] \qquad \text{(2.5a)} \\
=& ELBO(\mathbf{x}, y, d) - \mathbb{E}_{c\sim q_{\theta_c}(C|\mathbf{x})} \left[\log q_{\phi_s}(d \mid c) \right] - \mathbb{E}_{s\sim q_{\theta_s}(S|\mathbf{x})} \left[\log q_{\phi_c}(y \mid s) \right]. \qquad \text{(2.5b)}
\end{aligned}$$

In Equation (2.5a), the first two terms indicate the Kullback-Leibler divergence between the latent variables C and S and their prior distributions. In practice, we assume that the priors $p(C)$ and $p(S)$ follow standard multivariate Gaussian distributions. The third and fourth terms contain the approximated log-likelihoods of label predictions and the disentanglement

of the content and style. The last term stands for estimated distribution of $\mathbf{x}$. Note that in Equation (2.5b), our approximated $E\tilde{LBO}$ is composed of two parts: the original $ELBO$ which could be obtained from the factorization in Equation (2.1), and two regularization terms that aims to disentangle C and S through maximizing the log-likelihoods $\log q_{\phi_s}(d \mid c)$ and $\log q_{\phi_c}(y \mid s)$, which is shown by the dashed lines in Figure 2.2. By maximizing $E\tilde{LBO}$, we can train an accurate class predictor which is invariant to different styles. The detailed derivation is provided in supplementary material. Next, we introduce our data augmentation to assist in harnessing OOD examples.

2.3.2 Data Augmentation with Content and Style Intervention

After disentangling content and style, we try to harness OOD examples via two opposite augmentation procedures, namely *positive data augmentation* and *negative data augmentation* which aim to produce benign OOD data $\bar{\mathbf{x}}$ and malign OOD data $\hat{\mathbf{x}}$, respectively, so as to further enhance model generalization and improve robustness against anomalies. Specifically, to achieve this, positive data augmentation only conducts intervention on the style feature meanwhile keeping the content information the same; and the negative data augmentation attempts to affect the content feature while leaving the style unchanged, so as to produce malign OOD data, as shown in Figure 2.1 (b).

To achieve this goal, we employ adversarial data augmentation [96, 97, 98] which can directly conduct intervention on the latent variables without influencing each other, thus it is perfect for our intuition of augmenting content and style. Particularly, by adding a learnable perturbation $\mathbf{e}$ to each instance $\mathbf{x}$, we can obtain malign OOD data $\hat{\mathbf{x}}$ and benign OOD data $\bar{\mathbf{x}}$ with augmented content and style, respectively. For each data point $(\mathbf{x}, y, d)$, the perturbation $\mathbf{e}$ can be obtained through minimizing the intervention objective $\mathcal{L}(\cdot)$:

$$\mathbf{e} = \arg \min_{\mathbf{e};\|\mathbf{e}\|_p<\epsilon} \mathcal{L}(\mathbf{x} + \mathbf{e}, y, d; \theta_c, \phi_c, \theta_s, \phi_s), \tag{2.6}$$

where ϵ denotes the magnitude of the perturbation $\mathbf{e}$ with ℓ_p-norm. Since our goal of positive and negative data augmentation is completely different, here the intervention objective is designed differently for producing $\bar{\mathbf{x}}$ and $\hat{\mathbf{x}}$. For positive data augmentation, the intervention

objective is:

$$\mathcal{L}_{pos} = \mathcal{L}_d(g_c(\mathbf{x};\theta_c), g_c(\mathbf{x}+\mathbf{e};\theta_c)) - \mathcal{L}_{ce}(f_s(g_s(\mathbf{x}+\mathbf{e};\theta_s);\phi_s), d), \tag{2.7}$$

where the first term $\mathcal{L}_d(\cdot)$ indicates the distance measured between the contents extracted from the original instance and its perturbed version, and the second term $\mathcal{L}_{ce}(\cdot)$ denotes the cross-entropy loss. By minimizing $\mathcal{L}_{pos}$, the perturbation $\mathbf{e}$ would not significantly affect the content feature, meanwhile introducing a novel style that is distinct from its original domain d. Consequently, the augmented benign data with novel styles can be utilized to train a style-invariant model that is resistant to domain shift. Moreover, a specific style with domain label d' can be injected via modifying $\mathcal{L}_{pos}$ as:

$$\mathcal{L}'_{pos} = \mathcal{L}_d(g_c(\mathbf{x};\theta_c), g_c(\mathbf{x}+\mathbf{e};\theta_c)) + \mathcal{L}_{ce}(f_s(g_s(\mathbf{x}+\mathbf{e};\theta_s);\phi_s), d'). \tag{2.8}$$

Different from Equation (2.7), we hope to minimize the cross-entropy loss such that the perturbed instance can contain the style information from a target domain d'. As a result, the augmented benign data can successfully bridge the gap between source domain and target domain, and further improve the test performance in the target distribution.

As for negative data augmentation, the intervention objective is defined as:

$$\mathcal{L}_{neg} = \mathcal{L}_d(g_s(\mathbf{x};\theta_s), g_s(\mathbf{x}+\mathbf{e};\theta_s)) - \mathcal{L}_{ce}(f_c(g_c(\mathbf{x}+\mathbf{e};\theta_c);\phi_c), y). \tag{2.9}$$

By minimizing $\mathcal{L}_{neg}$, the perturbation would not greatly change the style information but would deviate the content from its original one with class label y. Subsequently, by recognizing the augmented malign data as unknown, the trained model would be robust to deceiving anomalies with familiar styles, thus boosting the OOD detection performance.

To accomplish the adversarial data augmentation process, here we perform multi-step projected gradient descent [99, 100]. Formally, the optimal $\bar{\mathbf{x}}$ and $\hat{\mathbf{x}}$ can be iteratively found:

$$\bar{\mathbf{x}}^{t+1} = \bar{\mathbf{x}}^t + \arg\min_{\mathbf{e}^t;\|\mathbf{e}^t\|_p<\epsilon} \mathcal{L}_{pos}(\bar{\mathbf{x}}^t + \mathbf{e}^t), \hat{\mathbf{x}}^{t+1} = \hat{\mathbf{x}}^t + \arg\min_{\mathbf{e}^t;\|\mathbf{e}^t\|_p<\epsilon} \mathcal{L}_{neg}(\hat{\mathbf{x}}^t + \mathbf{e}^t). \tag{2.10}$$

where the final iteration t is set to 15 in practice. Further, the optimal augmented data will be incorporated into model training, which is described in the next section.

Algorithm 1 Training process of HOOD

Labeled set $\mathcal{D}^l = \{(\mathbf{x}_i, y_i)\}_{i=1}^l$, unlabeled set $\mathcal{D}^u = \{(\mathbf{x}_i)\}_{i=1}^u$.
for $i = 1 \dots Max\_Iter$ **do**
  Pre-train the variational inference framework through maximizing $\tilde{ELBO}$ in Equation (2.5a);
  Assigning pseudo labels y^{ps} for unlabeled data $\mathcal{D}^u := \{(\mathbf{x}_i; y_i^{ps})\}_{i=1}^u$;
  if $i == Augmentation\_Iter$ **then**
    Conduct Adversarial Data Augmentation to obtain $\bar{\mathbf{x}}$ and $\hat{\mathbf{x}}$ via Equation (2.10);
    Add $\bar{\mathbf{x}}$ and $\hat{\mathbf{x}}$ into $\bar{\mathcal{D}}$ and $\hat{\mathcal{D}}$, respectively;
  end if
  Enumerate $\bar{\mathcal{D}}$ and conduct supervised training for each $\bar{\mathbf{x}}$;
  Enumerate $\hat{\mathcal{D}}$ and recognize each $\hat{\mathbf{x}}$ as unknown;
end for

2.3.3 Model Training with Benign and Malign OOD data

Finally, based on the above disentanglement and data augmentation in Section 2.3.1 and Section 2.3.2, we can obtain a benign OOD data $\bar{\mathbf{x}}$ and a malign OOD data $\hat{\mathbf{x}}$ from each data point $(\mathbf{x}, y, d)$, which will be appended to the benign dataset $\bar{\mathcal{D}}$ and malign dataset $\hat{\mathcal{D}}$, respectively. For utilization of benign OOD data $\bar{\mathbf{x}}$, we assign it with the original class label y and perform supervised training. For separation of malign OOD data $\hat{\mathbf{x}}$, we employ a one-vs-all classifier [79] to recognize them as unknown data that is distinct from its original class label y. The proposed HOOD method is summarized in Algorithm 1. Below, we specify the proposed HOOD algorothm to three typical applications with OOD data, namely OOD detection, open-set SSL, and open-set DA.

2.3.4 Deployment to OOD applications

Generally, in all three investigated applications, we are given a labeled set $\mathcal{D}^l = \{(\mathbf{x}_i, y_i)\}_{i=1}^l$ containing l labeled examples drawn from data distribution P^l, and an unlabeled set $\mathcal{D}^u = \{\mathbf{x}_i\}_{i=1}^u$ composed of u unlabeled examples sampled from data distribution P^u. Moreover, the label space of $\mathcal{D}^l$ and $\mathcal{D}^u$ are defined as $\mathcal{Y}^l$ and $\mathcal{Y}^u$, respectively.

OOD detection. The labeled set is used for training, and the unlabeled set is used as a test set which contains both ID data and malign OOD data. Particularly, the data distribution of

unlabeled ID data Q_{id} is the same as distribution P, but the distribution of OOD data P^u_{ood} is different from P, i.e., $P^u_{id} = P^l \neq P^u_{ood}$. The goal is to correctly distinguish OOD data from ID data in the test phase. During training, we conduct data augmentation to obtain domain label d, then follow our variational framework to update the model parameters. During test, we only use the content branch to predict the OOD score which is produced by the one-vs-all classifier. An instance is considered as an ID datum if the OOD score is smaller than 0.5, and an OOD datum otherwise.

Open-set SSL. The labeled set $\mathcal{D}^l$ and unlabeled set $\mathcal{D}^u$ are both used for training, and they are sampled from the same data distribution with different label spaces. Specifically, the unlabeled data contain some ID data that have the same classes as $\mathcal{D}^l$, and the rest unlabeled OOD data are from some unknown classes that do not exist in $\mathcal{D}^l$, formally, $\mathcal{Y}^l \subset \mathcal{Y}^u, \mathcal{Y}^u \setminus \mathcal{Y}^l \neq \varnothing$ and $P^l(\mathbf{x} \mid y) = P^u(\mathbf{x} \mid y), y \in \mathcal{Y}^l$. The goal is to properly leverage the labeled data and unlabeled ID data without being misled by malign OOD data, and correctly classify test data with labels in $\mathcal{Y}^l$. The training process is similar to OOD detection, except that HOOD would produce an OOD score for each unlabeled data. If an unlabeled instance is recognized as OOD data, it would be left out.

Open-set DA. The labeled set is drawn from source distribution P^l which is different from the target distribution P^u of unlabeled set. In addition, the label space $\mathcal{Y}^l$ is also a subset of $\mathcal{Y}^u$. Therefore, the unlabeled data consist of benign OOD data which have the same class labels as labeled data, and malign OOD data which have distinct data distribution as well as class labels from labeled data, formally, $P^l \neq P^u, \mathcal{Y}^l \subset \mathcal{Y}^u, \mathcal{Y}^u \setminus \mathcal{Y}^l \neq \varnothing$. The goal is to transfer the knowledge of labeled data to the benign OOD data, meanwhile identify the malign OOD data as unknown. In this application, we assign each target instance with a domain label to distinguish them from other augmented data. Then we alter the positive data augmentation objective from Equation (2.7) to Equation (2.8). During test, HOOD would predict each target instance as some class if it is benign, and as unknown otherwise.

2.4 Experiment

In this section, we first describe the implementation details. Then, we experimentally validate our method on three applications, namely OOD detection, open-set SSL, and open-set DA. Finally, we present extensive performance analysis on our disentanglement and intervention modules. Additional details and quantitative findings can be found in the supplementary material.

2.4.1 Implementation Details

In experiments, we choose Wide ResNet-28-2 [101] for OOD detection and Open-set SSL tasks, and follow [102, 103] to utilize ResNet50 pre-trained on Imagenet [104] for Open-set DA. For implementing HOOD, we randomly choose 4 augmentation methods from the transformation pool in RandAugment [105], to simulate different styles. The pre-training iteration $Augmentation\_Iter$ is set to 100,000, and the perturbation magnitude $\epsilon = 0.03$, following [98] in all experiments. Next, we validate HOOD in three applications.

2.4.2 OOD detection

In OOD detection task, we use SVHN [106] and CIFAR10 [107] as the ID datasets, and use LSUN [108], DTD [109], CUB [110], Flowers [111], Caltech [112], and Dogs [113] datasets as the OOD datasets that occur during test phase. Particularly, to explore the model generalization ability, we only sample 100 labeled data and 20,000 unlabeled data from each class and conduct semi-supervised training, then we test the trained model on the unlabeled OOD dataset. To evaluate the performance, we utilize AUROC [28] which is an essential metric for OOD detection, and a higher AUROC value indicates a better performance.

For comparison, we choose some typical OOD detection methods including Likelihood [28] which simply utilizes softmax score as the detection criterion, ODIN [35] which enhances the performance of Likelihood through adding adversarial attack, Likelihood Ratio [74] which modifies the softmax score through focusing on the semantic feature, and OpenGAN [58]

TABLE 2.1. Comparison with typical OOD detections methods. Averaged AUROC (%) with standard deviations are computed over three independent trails. The best results are highlighted in bold.

OOD dataset	LSUN	DTD	CUB	Flowers	Caltech	Dogs
ID dataset			SVHN			
Likelihood	52.25 ± 0.3	50.33 ± 0.7	48.76 ± 0.6	47.33 ± 0.2	51.54 ± 0.4	54.34 ± 0.4
ODIN	55.72 ± 0.2	53.32 ± 0.5	52.70 ± 0.4	50.47 ± 0.7	56.41 ± 0.4	61.16 ± 0.3
Likelihood Ratio	79.34 ± 0.5	78.42 ± 0.3	75.90 ± 0.7	74.53 ± 0.4	76.25 ± 0.3	83.55 ± 0.4
OpenGAN	83.77 ± 0.4	80.36 ± 0.5	77.49 ± 0.8	79.26 ± 0.5	$\mathbf{86.66 \pm 0.5}$	86.84 ± 0.5
HOOD	$\mathbf{84.10 \pm 0.6}$	$\mathbf{80.68 \pm 0.6}$	$\mathbf{79.24 \pm 0.5}$	$\mathbf{80.93 \pm 0.7}$	85.34 ± 0.7	$\mathbf{87.58 \pm 0.8}$
ID dataset			CIFAR10			
Likelihood	54.32 ± 0.5	52.16 ± 0.4	50.67 ± 0.4	49.26 ± 0.3	53.86 ± 0.4	56.92 ± 0.2
ODIN	58.60 ± 0.3	55.59 ± 0.6	58.48 ± 0.7	51.44 ± 0.9	59.36 ± 0.4	64.82 ± 0.5
Likelihood Ratio	81.41 ± 0.6	79.77 ± 0.5	79.35 ± 0.8	77.17 ± 0.7	80.67 ± 0.5	86.76 ± 0.3
OpenGAN	84.03 ± 0.4	81.29 ± 0.8	82.84 ± 1.0	$\mathbf{82.32 \pm 0.4}$	86.78 ± 0.3	90.14 ± 0.5
HOOD	$\mathbf{86.12 \pm 0.6}$	$\mathbf{83.64 \pm 0.5}$	$\mathbf{83.53 \pm 0.6}$	81.56 ± 0.8	$\mathbf{87.24 \pm 0.8}$	$\mathbf{90.86 \pm 0.6}$

which can further improve the performance via separating the generated "hard" examples that are deceivingly close to ID data.

The experimental results are shown in Table 2.1, we can see that HOOD can greatly surpass Likelihood, ODIN, and Likelihood Ratio, and can outperform OpenGAN in most scenarios. When compared with softmax-prediction-based methods such as Likelihood and ODIN, HOOD surpasses them in a large margin, as HOOD can correctly separate some overconfident OOD examples from ID data. As for Likelihood Ratio, our method can achieve better performance through producing "hard" malign OOD data, thus successfully avoiding deceiving examples that are extremely close to ID data. Although both OpenGAN and HOOD generate "hard" malign data to train an open classifier, HOOD can successfully distinguish content and style thanks to the aforementioned disentanglement, thus avoid rejecting too much benign OOD data and further yield better detection performance than OpenGAN.

2.4.3 Open-Set SSL

In open-set SSL task, we follow [70] to construct our training dataset using two benchmark datasets CIFAR10 and CIFAR100 [107], which contains 10 and 100 classes, respectively. The constructed dataset has 20,000 randomly sampled unlabeled data and a varied number

TABLE 2.2. Comparison with typical Open-set SSL methods. Averaged test accuracies (%) with standard deviations are computed over three independent trails. The best results are highlighted in bold.

Training dataset		CIFAR10			CIFAR100		
No. of Labeled data		50	100	400	50	100	400
Clean Acc.	UASD	72.82 ± 0.9	75.53 ± 1.8	76.74 ± 1.7	58.87 ± 0.6	61.68 ± 1.2	65.97 ± 2.4
	DS3L	74.44 ± 1.3	76.89 ± 1.5	78.80 ± 0.6	60.40 ± 0.5	64.35 ± 1.5	67.65 ± 1.3
	MTCF	79.88 ± 1.3	81.41 ± 1.0	83.92 ± 0.8	62.78 ± 0.5	65.84 ± 2.1	69.46 ± 0.6
	OpenMatch	$\mathbf{84.10 \pm 1.1}$	$\mathbf{85.30 \pm 0.4}$	$\mathbf{87.92 \pm 1.0}$	65.76 ± 0.9	$\mathbf{68.46 \pm 0.5}$	72.87 ± 1.4
	T2T	82.74 ± 1.2	83.56 ± 1.4	85.97 ± 0.8	65.16 ± 1.2	67.58 ± 0.9	71.96 ± 1.1
	HOOD	83.55 ± 1.2	84.16 ± 1.5	86.22 ± 2.7	$\mathbf{66.39 \pm 1.7}$	68.03 ± 2.6	$\mathbf{73.32 \pm 0.6}$
Corrupted Acc.	UASD	39.36 ± 1.2	41.38 ± 0.7	42.66 ± 1.8	31.55 ± 2.0	33.39 ± 1.7	35.20 ± 0.8
	DS3L	39.97 ± 0.8	42.58 ± 0.8	44.39 ± 0.6	33.72 ± 0.8	34.67 ± 0.8	36.64 ± 0.6
	MTCF	40.16 ± 1.2	40.58 ± 1.1	43.33 ± 0.7	32.72 ± 0.8	34.33 ± 2.3	35.53 ± 0.6
	OpenMatch	41.38 ± 0.7	42.90 ± 0.6	45.79 ± 0.8	35.98 ± 1.3	36.47 ± 0.7	38.56 ± 0.6
	T2T	41.39 ± 1.6	45.56 ± 1.6	49.88 ± 1.5	$\mathbf{41.03 \pm 1.7}$	39.64 ± 0.7	41.38 ± 1.6
	HOOD	$\mathbf{44.42 \pm 1.7}$	$\mathbf{48.38 \pm 0.9}$	$\mathbf{50.74 \pm 0.6}$	40.82 ± 1.5	$\mathbf{41.65 \pm 0.9}$	$\mathbf{43.72 \pm 2.2}$

of labeled data. Here the number of labeled data is set to 50, 100, and 400 per class in both CIFAR10 and CIFAR100. Moreover, to create the open-set problem in CIFAR10, the unlabeled data is sampled from all 10 classes and the labeled data is sampled from the 6 animal classes. As for CIFAR100, the unlabeled data are sampled from all 100 classes and the labeled data is sampled from the first 60 classes. For evaluation, we first use the test dataset from the original CIFAR10 and CIFAR100 and denote the test accuracy as "Clean Acc.". Further, to evaluate the capability of handling OOD examples, we test on CIFAR10-C and CIFAR100-C [114] which add different types of corruptions to CIFAR10 and CIFAR100, respectively. The test accuracy from the corrupted datasets can reveal the robustness of neural networks against corruptions and perturbations, and it is denoted as "Corrupted Acc.".

For comparison, we choose some typical open-set SSL methods including Uncertainty-Aware Self-Distillation method UASD [71] and T2T [115] which filters out the OOD data via using OOD detection, Safe Deep Semi-Supervised Learning DS3L [70] which employs meta-learning to down-weight the OOD data, Multi-Task Curriculum Framework MTCF [49] which recognizes the OOD data as different domain, and OpenMatch [48] which utilizes open-set consistency training on OOD data.

TABLE 2.3. Comparison with typical Open-set DA methods. Averaged test accuracies (%) with standard deviations are computed over three independent trails. The best results are highlighted in bold.

Dataset	Office						VisDA
Domain	A→W	A→D	D→W	W→D	D→A	W→A	Synthetic→Real
OSBP	86.5 ± 2.0	88.6 ± 1.4	97.0 ± 1.0	97.9 ± 0.9	88.9 ± 2.5	85.8 ± 2.5	62.9 ± 1.3
UAN	87.7 ± 1.2	87.0 ± 0.8	93.5 ± 1.3	97.2 ± 1.6	88.4 ± 0.7	87.8 ± 1.6	63.8 ± 2.4
STA	89.5 ± 0.6	93.7 ± 1.5	97.5 ± 0.2	$\mathbf{99.5 \pm 0.2}$	89.1 ± 0.5	87.9 ± 0.9	66.4 ± 1.3
HOOD	$\mathbf{90.1 \pm 1.5}$	$\mathbf{94.2 \pm 1.4}$	$\mathbf{99.6 \pm 0.6}$	98.3 ± 0.9	$\mathbf{89.8 \pm 0.8}$	$\mathbf{91.3 \pm 1.8}$	$\mathbf{72.4 \pm 1.6}$

The experimental results are shown in Table 2.2. Compared to the strongest baseline method OpenMatch, which randomly samples eleven different transformations from a transformation pool, our method has transformations that are limited to only four types. In CIFAR10 and CIFAR100 regarding the Clean Acc., the proposed HOOD is slightly outperformed by OpenMatch. However, thanks to the disentanglement, HOOD can be invariant to different styles and focus on the content feature. Therefore, when facing corruption, HOOD can be more robust than all baseline methods. As shown by the Corrupted Acc. results, our method surpasses OpenMatch for more than 3%.

2.4.4 Open-Set DA

In open-set DA task, we follow [41] to validate on two DA benchmark datasets Office [116] and VisDA [117]. Office dataset contains three domains Amazon (A), Webcam (W), and DSLR (D), and each domain is composed of 31 classes. VisDA dataset contains two domains Sythetic and Real, and each domain consists of 12 classes. To create an open-set situation in Office, we follow [41, 40] to construct the source dataset by sampling from the first 21 classes in alphabetical order. Then, the target dataset is sampled from all 31 classes. As for VisDA, we choose the first 6 classes for source domain, and use all the 12 classes for target domain. We use "A→W" to indicate the transfer from "A" domain to "W" domain.

For comparison, we choose three typical open-set DA approaches including Open-Set DA by BackPropagation OSBP [41] which employs an OpenMax classifier to recognize unknown classes and perform gradient flipping for open-set DA, Universal Adaptation Network

TABLE 2.4. Ablation study on necessity of each module.

Application	OOD detection	Open-Set SSL	Open-Set DA
w/o disentanglement	84.94 ± 1.3	82.55 ± 1.1	64.6 ± 0.9
w/o benign OOD data	85.95 ± 1.8	83.32 ± 2.0	66.3 ± 2.5
w/o malign OOD data	82.50 ± 2.2	85.40 ± 0.8	71.8 ± 1.2
w/o both augmentations	80.83 ± 0.8	81.14 ± 1.2	65.4 ± 1.2
HOOD	$\mathbf{86.12 \pm 0.6}$	$\mathbf{86.22 \pm 2.7}$	$\mathbf{72.4 \pm 1.6}$

UAN [102] which utilize entropy and domain similarity to down-weight malign OOD data, and Separate To Adapt STA [40] which utilizes SVM to separate the malign OOD data.

The experimental results are shown in Table 2.3. Compared to the baseline methods, the proposed HOOD is largely benefited from the generated benign OOD data, which have two major strengths: (1) they resemble target domain data by having common styles, and (2) their labels are accessible as they share the same content as their corresponding source data. Therefore, through conducting supervised training such benign OOD data, the domain gap can be further mitigated, thus achieving better performance than baseline methods. Quantitative results show that HOOD can surpass other methods in most scenarios. Especially in VisDA, HOOD can outperform the second-best method with 6% improvement, which proves the effectiveness of HOOD in dealing with open-set DA.

2.4.5 Performance Analysis

Ablation Study: To verify the effectiveness of each module, we conduct an ablation study on three OOD applications by eliminating one component at a time. Specifically, our HOOD can be ablated into: "w/o disentanglement" which indicates removing the disentanglement loss in Equation (2.5a), "w/o benign OOD data" which denotes training without benign OOD data, "w/o malign OOD data" which stands for discarding malign OOD data, and "w/o both augmentations" indicates training without both benign and malign OOD data. In OOD detection, we use CIFAR10 and LSUN as the ID and OOD dataset, respectively. In open-set SSL, we choose CIFAR10 with 400 labels for each class. As for open-set DA, we use VisDA dataset.

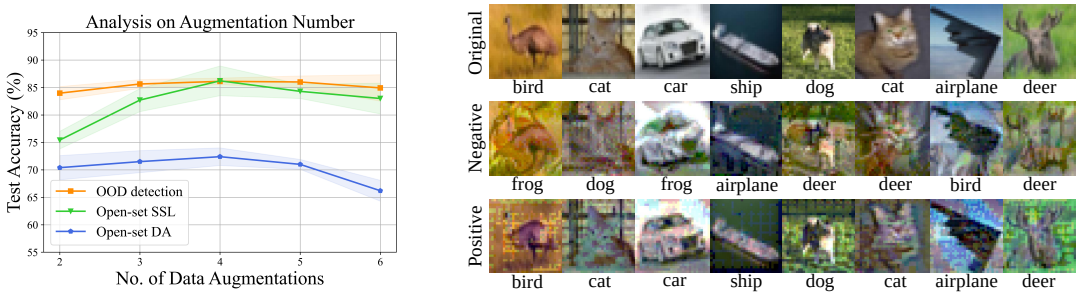


FIGURE 2.3. Left: Augmentation number analysis. Right: CIFAR10 Visualization of our data augmentation.

The experimental results are shown in Table 2.4. We can see each module influences the performance differently in three applications. First, we can see that the malign OOD data is essential for OOD detection, as it can act as unknown anomalies and reduce the overconfidence in unseen data. Then, benign OOD data can largely improve the learning performance in open-set SSL and open-set DA, as they can enforce the model to focus on the content feature for classification. Additionally, we can see that discarding both benign and malign OOD data shows performance degradation compared to both "w/o benign OOD data" and "w/o malign OOD data". Therefore, our HOOD can correctly change the style and content, which can correspondingly benefit generalization tasks (such as open-set DA and open-set SSL) and detection tasks (such as OOD detection). Moreover, open-set DA relies more on the disentanglement than the rest two modules, owing to the disentanglement can exclude the style changing across different domains. Hence, our disentanglement can effectively eliminate the distribution change from different domains and help learn invariant features.

Analysis on Augmentation Number: Since HOOD does not introduce any hyper-parameter, the most influential setting is the number of data augmentation. To analyze its influence on the learning results, we vary the number of augmentations that are sampled from the RandAugment Pool [105] from 2 to 6. The results are shown in Figure 2.3 left. We can see that both too less and too many augmentations would hurt the results. This is because a small augmentation number would undermine the generalization to various styles; and a large augmentation number would increase the classification difficulty of the style branch, further making the disentanglement hard to achieve. Therefore, setting the augmentation number to 4 is reasonable.

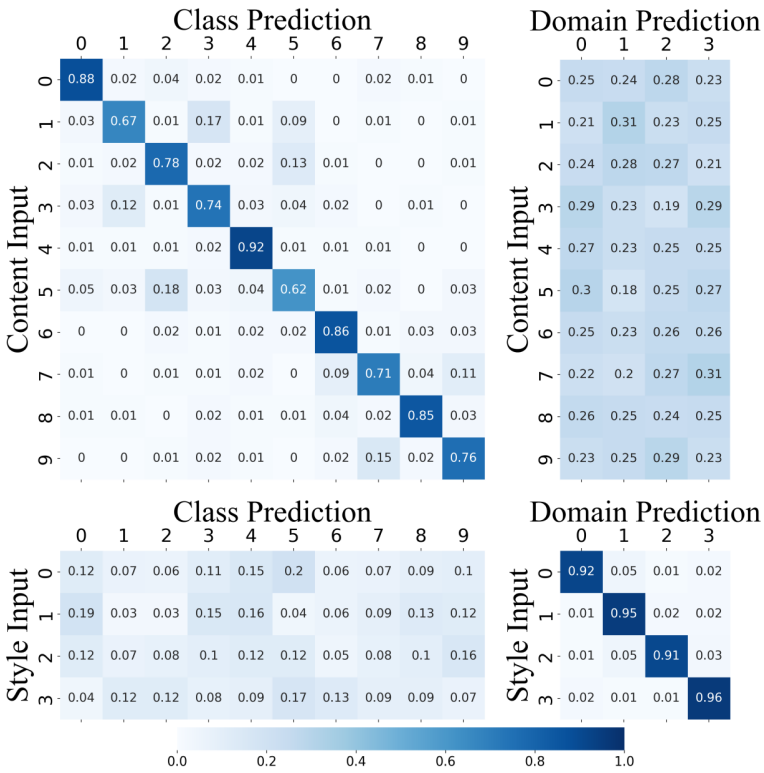


FIGURE 2.4. Illustration content ande style disentanglement on CIFAR10. The number in each cell denotes the prediction probability.

Visualization: Furthermore, to show the effect of our data augmentations, we visualize the augmented images by applying large perturbation magnitude (4.7) [118] in Figure 2.3 right. The model prediction is shown below each image. We can see that the negative data augmentation significantly changes the content which is almost unidentifiable. However, positive data augmentation can still preserve most of the content information and only change the style of images. Therefore, the augmented data are tailor-designed for training a robust classifier.

Content and Style Disentanglement: To further testify that our disentanglement between content and style is effective, we select the latent variables from different content and style categories, and use the learned class and domain classifiers for cross-prediction. Specifically, there are four kinds of input-prediction types: content-class, content-domain, style-class, and style-domain. As we can see in Figure 2.4, only the content features are meaningful for class prediction, and the same phenomenon goes for style input and domain prediction. However,

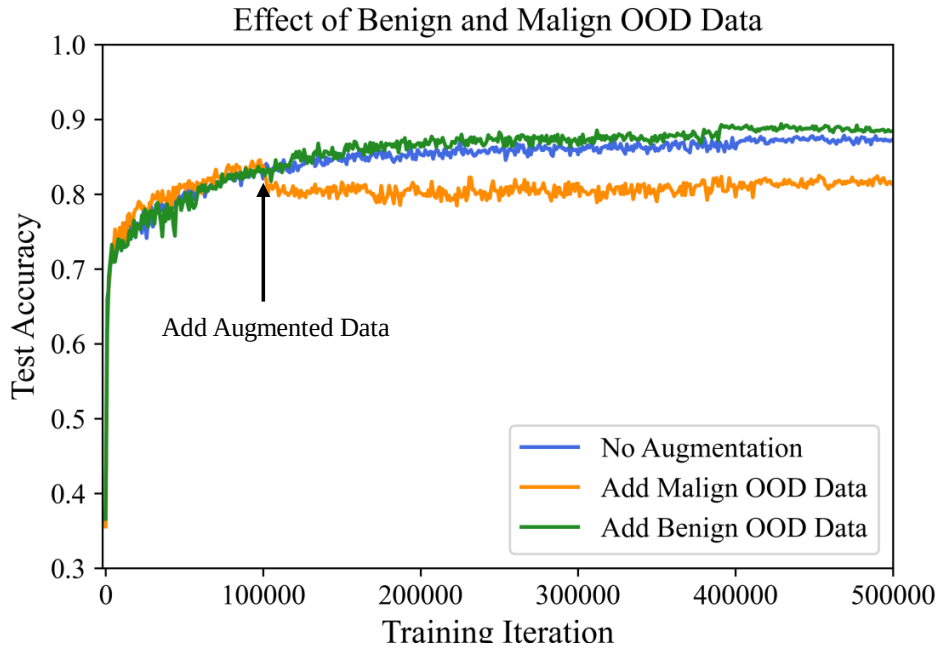


FIGURE 2.5. Effect of adding benign and malign OOD data into training.

neither of the style and content features can be identified by the class predictor and domain predictor, respectively. Therefore, we can reasonably conclude that our disentanglement between content and style is effectively achieved.

Effect of Adding Benign and Malign OOD Data into Training: To give a illustrative comparison of adding benign OOD data and malign OOD data into training, we conduct experiments under the open-set SSL setting and separately augmenting benign OOD data and malign OOD data to compare their effects. Moreover, we conduct plain training as a baseline result which do not use either augmentations. The results are shown in Figure 2.5. We can see that after adding augmented data, the effect of malign OOD data causes sudden performance degradation. On the contrary, benign OOD data can further improve the learning result compared to the plain training baseline. Which again shows that preserving content and augmenting style is beneficial, and eliminating content is harmful for generalization.

Analysis of OOD Score To show the effectiveness of identifying malign OOD data from benign OOD data, we test the performance of HOOD on three applications to observe the OOD scores of benign OOD data and malign OOD data and show the averaged OOD scores

TABLE 2.5. Averaged OOD scores on three applications.

Application	OOD score	
	Benign OOD data	Malign OOD data
OOD detection	0.16 ± 0.3	0.83 ± 0.6
Open-Set SSL	0.08 ± 0.5	0.91 ± 0.4
Open-Set DA	0.21 ± 0.4	0.88 ± 0.3

TABLE 2.6. Execution efficiency comparisons on three applications.

OOD Detection		Open-set SSL		Open-set DA	
Method	Time	Method	Time	Method	Time
Likelihood	6.2h	DS3L	15.4h	UAN	8.5h
OpenGAN	7.8h	OpenMatch	10.5h	STA	9.1h
HOOD	11.4h	HOOD	13.7h	HOOD	12.4h

in Table 2.5. We can see that the OOD score produced by our one-vs-all classifier can clearly distinguish benign and malign OOD data during the test phase, which again validates the effectiveness of HOOD.

Execution Efficiency Additionally, to give a quantitative comparison on the execution efficiency of HOOD, here we provide the running time on 3090 GPU compared to some typical baseline methods. The results are shown in Table 2.6. Note that our method involves causal disentanglement as well as adversarial training, therefore, the training time is increased.

2.5 Conclusion

In this Chatper, we propose HOOD to effectively harness OOD examples. Specifically, we construct a SCM to disentangle content and style, which can be leveraged to identify benign and malign OOD data. Subsequently, by maximizing ELBO, we can successfully disentangle the content and style feature and break the spurious correlation between class and domain. As a result, HOOD can be more robust when facing distribution shifts and unseen OOD data. Furthermore, we augment the content and style through a novel intervention process to produce benign and malign OOD data, which can be leveraged to improve classification and

OOD detection performance. Extensive experiments are conducted to empirically validate the effectiveness of HOOD on three typical OOD applications.

CHAPTER 3

Sharpness-Based Distribution Robust Optimization

Robust generalization aims to tackle the most challenging data distributions, which are rare in the training set and contain severe noise, i.e., photon-limited corruptions. Common solutions, such as distributionally robust optimization (DRO), focus on the worst-case empirical risk to ensure low training error on the uncommon noisy distributions. However, due to the over-parameterized model being optimized on scarce worst-case data, DRO fails to produce a smooth loss landscape, thus struggling on generalizing well to the test set. Therefore, instead of focusing on the worst-case risk minimization, we propose SharpDRO by penalizing the sharpness of the worst-case distribution, which measures the loss changes around the neighborhood of learning parameters. Through worst-case sharpness minimization, the proposed method successfully produces a flat loss curve on the corrupted distributions, thus achieving robust generalization. Moreover, by considering whether the distribution annotation is available, we apply SharpDRO to two problem settings and design a worst-case selection process for robust generalization. Theoretically, we demonstrate that SharpDRO offers a strong convergence guarantee. Experimentally, we simulate photon-limited corruptions using the CIFAR10/100 and ImageNet30 datasets and demonstrate that SharpDRO exhibits a strong generalization ability against severe corruptions, outperforming well-known baseline methods by a significant margin.

3.1 Introduction

Learning against corruptions has been a vital challenge in the practical deployment of computer vision models, as learning models are much more fragile to subtle noises than human

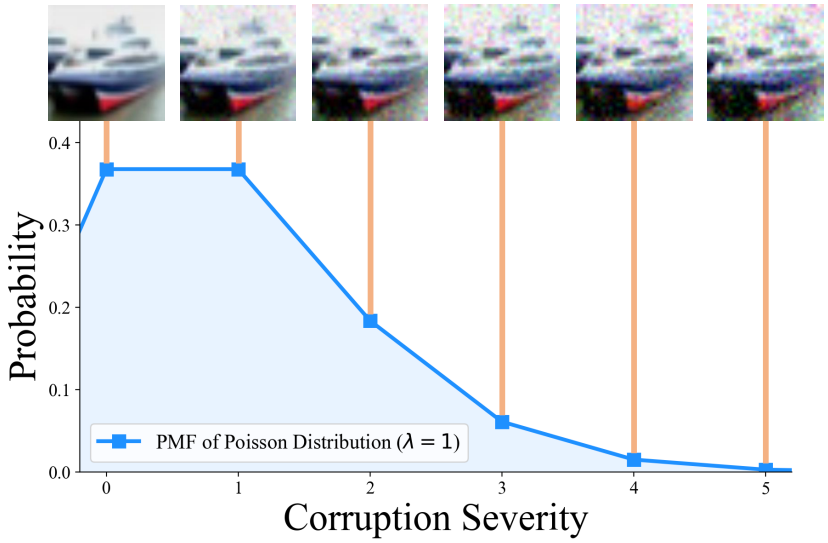


FIGURE 3.1. Illustration of photon-limited corruptions.

perception systems [119, 28, 120]. During the training, the encountered corruptions are essentially perceived as a distribution shift, which would significantly hinder the prediction results [35, 121, 31, 122, 123, 124]. Therefore, to mitigate the performance degradation, enhancing generalization to corrupted data distributions has drawn lots of attention [69, 125].

In the real world, noise corruptions are commonly known as photon-limited imaging problems [126, 127, 128, 129] which arise due to numerous small corruption photons arriving at an image sensor. Consequently, different numbers of captured photons would form different levels of corruption severity, further producing multiple data distributions and imposing varied impacts on learning models [114]. Specifically, the encountered photon-limited corruption $\mathscr{E}$ is a composition of multiple noise photons u, which is triggered by some discrete factors with a certain probability during a time interval. For example, a photon u can be triggered by each platform changing, redistribution, transmission, etc. More photons are captured, and severe corruption would be applied to the image. Therefore, the severity s of the photon-limited corruptions $\mathscr{E}$ can be modeled by Poisson distribution, i.e., $s \sim P(s;\lambda) = \frac{e^{-\lambda}\lambda^s}{s!}$, which is illustrated in Figure 3.1. As a result, the real-world dataset is not completely composed of clean data, but contains corrupted data with various severities.

Dealing with such a realistic problem by vanilla empirical risk minimization can achieve satisfactory average accuracy on the whole training set. However, due to the extremely limited

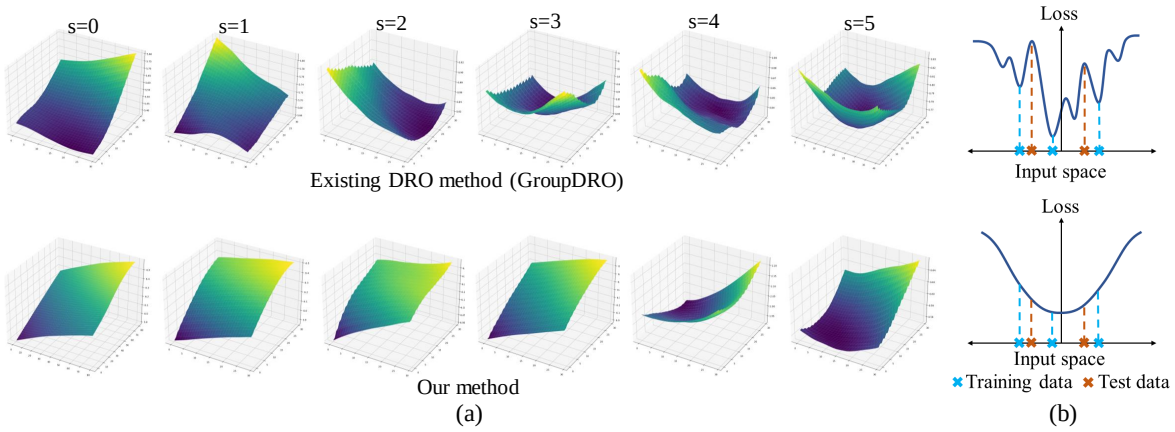


FIGURE 3.2. Illustration of our motivation. (a) Loss surface visualization of GroupDRO and the proposed SharpDRO. The columns from left to right stand for corrupted distributions with severity $s = 0$ to 5. (b) Illustration of why a sharp loss surface hinders generalization to test data.

number of severely corrupted data, the learning model would produce large training errors on the corrupted distributions, further hindering the robust performance under challenging real-world situations. A popular approach to achieve low error on the scarce corrupted data is distributionally robust optimization (DRO) [130, 125, 131, 132, 133, 134], which commonly optimizes the model parameter θ by optimizing:

$$\min_{\theta \in \Theta} \sup_{Q \in \mathcal{Q}} \mathbb{E}_{(x,y)\sim Q} \left[\mathcal{L}(\theta; (x, y)) \right], \tag{3.1}$$

where $\mathcal{Q}$ denotes the uncertainty set that is utilized to estimate the possible test distribution. Intuitively, DRO assumes that $\mathcal{Q}$ consists of multiple sub-distributions, among which exists a worst-case distribution Q. By concentrating on the risk minimization of the worst-case distribution, DRO hopes to train a robust model that can deal with the potential distribution shift during the test phase. However, existing DRO methods usually leverage over-parameterized models to focus on a small portion of worst-case training data. Therefore, the worst-case data contaminated with severe corruption is highly possible to get stuck in sharp minima. As shown in the upper part of Figure 3.2 (a), a stronger corruption would cause the existing method to learn a sharper loss surface. Consequently, optimization via DRO fails to produce a flat loss landscape over the corrupted distributions, which leads to a large generalization gap between training and test set [135, 136].

To remedy this defect, in this Chapter, we propose the SharpDRO method to focus on learning a flat loss landscape of the worst-case data, which can largely mitigate the training-test generalization gap problem of DRO. Specifically, we adopt the largest loss difference formed by applying weight perturbation [137, 138] to measure the sharpness of the loss function. Intuitively, a sharp loss landscape is sensitive to noise and cannot generalize well on the test set. On the contrary, a flat loss landscape produces consistent loss values and is robust against perturbations (Figure 3.2 (b)). By minimizing the sharpness, we can effectively enhance the generalization performance [135, 136]. However, directly applying sharpness minimization on multiple distributions would yield poor results [139], as the computed sharpness could be influenced by the largest data distribution, and thus cannot generalize well to small corrupted data. Therefore, we only focus on worst-case sharpness minimization. In this way, as the lower part of Figure 3.2 (a) shows, SharpDRO successfully produces a flat loss surface, thus achieving robust generalization on the severely corrupted distributions.

In addition, identification of the worst-case distribution requires expensive annotations, which are not always practically feasible [140]. In this Chapter, we apply SharpDRO to solve two problem settings: (1) *Distribution-aware robust generalization*, which assumes that distribution indices are accessible, and (2) *Distribution-agnostic robust generalization*, where the distributions are no longer identifiable, making the worst-case data hard to find. Existing approaches, such as Just Train Twice (JTT), require two-stage training, which is rather inconvenient. To tackle this challenge, we propose a simple (Out-of-distribution) OOD detection [28, 42, 43, 35, 36, 39, 141] process to detect the worst-case data, which can be further leveraged to enable worst-case sharpness minimization. Through constructing training sets according to the Poisson distributed noisy distribution using CIFAR10/100 and ImageNet30, we show that SharpDRO can achieve robust generalization results on both problem settings, surpassing well-known baseline methods by a large margin.

To sum up, our main contributions are threefold:

- We proposed a sharpness-based DRO method that overcomes the poor worst-case generalization performance of distributionally robust optimization.

- We apply our SharpDRO to both distribution-aware and distribution-agnostic settings, which brings a practical capability to our method. Moreover, we propose an OOD detection approach to select worst-case data to enable robust generalization.
- Theoretically, we show that SharpDRO has a convergence rate of $\mathcal{O}(\frac{\kappa^2}{\sqrt{MT}})$. Empirically, we form a photon-limited corruption dataset that follows a Poisson distribution, and conduct extensive experiments to show a strong generalization ability of SharpDRO as well as its superiority to compared baseline methods.

In the following, we first briefly introduce the background and discuss the problem setting in section 3.2. Then, we specify our SharpDRO over two problem settings in Section 3.3. Moreover, we give a detailed optimization process and provide convergence analysis in Section 3.3.3. Further, we conduct extensive experiments to validate our SharpDRO in Section 3.4. At last, we conclude this Chapter in Section 3.6.

3.2 Robust Generalization Methods

Due to the practical significance of robust generalization, various approaches have been proposed to deal with distribution shift. Here, we briefly introduce three typical baseline methods, namely Invariant Risk Minimization, Risk Extrapolation, and GroupDRO.

Invariant Risk Minimization (IRM) [69, 142, 143] aims to extract the invariant feature across different distributions (also denoted as environments). Specifically, the learning model is separated into a feature extractor G and a classifier C. IRM assumes an invariant model $C \circ G$ over various environments can be achieved if the classifier C constantly stays optimal. Then, the learning objective is formulated as:

$$\begin{aligned} &\min_{C^* \circ G} \left\{ \mathcal{L}_{\text{IRM}} := \sum_{e \in \mathcal{E}} \mathcal{L}^e(C^* \circ G) \right\} \\ &\text{s. t. } C^* \in \arg\min_{G} \mathcal{L}^e(C \circ G), \text{for all } e \in \mathcal{E}, \end{aligned} \tag{3.2}$$

where C^* stands for the optimal classifier, and e denotes a specific environment from a given environmental set $\mathcal{E}$. By solving Equation (3.2), the feature extractor G can successfully

learn invariant information without being influenced by the distribution shift between different environments.

Risk Extrapolation (REx) [144] targets at generalization to out-of-distribution (OOD) environments. Inspired by the discovery that penalizing the loss variance across distributions helps achieve improved performance on OOD generalization, REx proposes to optimize via:

$$\min_{\theta\in\Theta}\left\{\mathcal{L}_{\text{REx}} := \sum_{e\in\mathcal{E}}\mathcal{L}^e(\theta) + \beta Var(\mathcal{L}^e, ..., \mathcal{L}^m)\right\}, \tag{3.3}$$

where β controls the penalization level. Intuitively, REx seeks to achieve risk fairness among all m training environments, so as to increase the similarity of different learning tasks. As a result, the training model can capture the invariant information that helps generalize to unseen distributions.

GroupDRO [125, 145, 132] deal with the situation when the correlation between class label y and unknown attribute a differs in the training and test set. Such a difference is called spurious correlation, which could seriously misguide the model prediction. As a solution, GroupDRO considers each combination of class and attribute as a group g. By conducting risk minimization through:

$$\min_{\theta\in\Theta}\left\{\mathcal{L}_{\text{GroupDRO}} := \max_{g}\mathbb{E}_{(x,y)\sim P_g}\left[\mathcal{L}(\theta;(x,y))\right]\right\}. \tag{3.4}$$

The worst-case group from the distribution P_g, which commonly holds spurious correlation, is emphasized, thus breaking the spurious correlation.

Discussion: IRM and REx both focus on learning invariant knowledge across various environments. However, when the training set contains extremely imbalanced noisy distributions, as shown in Figure 3.1, the invariant learning results would be greatly misled by the most dominating distribution. Thus, the extracted invariant feature would be questionable for generalization against distribution shift. Although emphasizing the risk minimization of worst-case data via GroupDRO can alleviate the imbalance problem, its generalization performance is still sub-optimal when facing novel test data. However, SharpDRO can not only focus on the uncommon corrupted data, such as adversarial perturbation [146], but also

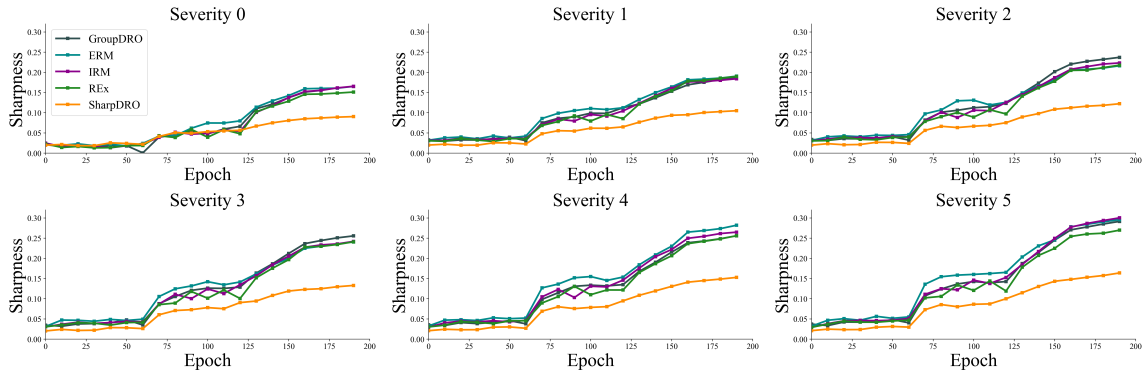


FIGURE 3.3. Sharpness during networking training on clean ($s = 0$) and corrupted distributions ($s = 1$ to 5).

effectively improve the generalization performance on the test set by leveraging worst-case sharpness minimization.

Our investigated problem is closely related to OOD generalization which is a broad field that contains many popular research topics, such as Domain Generalization [147, 148, 149, 150, 151, 152, 153, 154], Causal Invariant Learning [69, 144, 155, 156, 157, 158]. Generally, existing works mainly study two types of research problems: (1) mitigating domain shift between the training and test datasets; and (2) breaking the spurious correlation between causal factors. However, as generalization against corruptions **does not introduce any domain shift or spurious correlation**, such a problem cannot be naively solved by domain generalization methods or causal representation learning techniques. Therefore, in this Chapter, we focus on complementing this rarely-explored field and propose SharpDRO to enforce robust generalization against corruption. In the next section, we elaborate on the methodology of SharpDRO.

3.3 Methodology

In robust generalization problems, we are given a training set $\mathcal{D}^{\text{train}}$ containing n image examples, each example $x \in \mathcal{X}$ is given a class label $y \in \mathcal{Y} = \{1, 2, ..., c\}$. Moreover, the training set is corrupted by a certain type of noise whose severity s follows a Poisson distribution $P(s; \lambda)$. Here we assume $\lambda = 1$, which indicates that the mean number of noise

photons u that occurred during a time interval is 1. Therefore, the distribution P of the whole training set is composed of S sub-distributions P_s, $s \in \{1, 2, ..., S\}$ with varied levels of corruption. Our goal is to learn a robust model $\theta \in \Theta$ that can achieve good generalization performance on challenging data distributions P_s with large severity.

The general objective of SharpDRO is formulated as:

$$\min_{\theta} \left\{ \mathcal{L}_{\text{SharpDRO}} := \mathbb{E}_{(x,y)\sim Q} \left[\mathcal{L}(\theta; (x, y)) \right] + \mathbb{E}_{(x,y)\sim Q} \left[\mathcal{R}(\theta; (x, y)) \right] \right\}, \tag{3.5}$$

where the first term denotes the risk minimization using the loss function $\mathcal{L}$, meanwhile, a worst-case distribution Q is selected based on the model prediction. The second term $\mathcal{R}$ indicates the sharpness minimization, which aims to maximally improve the generalization performance on the worst-case distribution Q. Specifically, as shown in Figure 3.3, the sharpness gradually increases as the corruption severity increases. Therefore, to accomplish robust generalization, we are motivated to emphasize the worst-case distribution. As a result, we can produce much smaller sharpness compared to other methods, especially on severely corrupted distributions.

In the following, we first introduce worst-case sharpness for robust generalization. Then, we demonstrate worst-case data selection on two scenarios. Finally, we provide a detailed optimization process and convergence analysis.

3.3.1 Sharpness for Robust Generalization

The main challenge of robust generalization is that the training distribution is extremely imbalanced, as shown in Figure 3.1. The performance on the abundant clean data is quite satisfactory, but robustness regarding the corrupted distribution is highly limited, owing to the severe disturbance of corruption as well as the insufficiency of noisy data. To enhance the generalization performance, we leverage sharpness to fully exploit the worst-case data. Specifically, sharpness [137, 159, 160, 138, 161] is measured by the largest loss change

when model weight θ is perturbed with ϵ, formally:

$$\mathcal{R} := \max_{\|\epsilon\|_2 \leq \rho} \{\mathcal{L}(\theta + \epsilon; (x, y)) - \mathcal{L}(\theta; (x, y))\}, \tag{3.6}$$

where ρ is a scale parameter to control the perturbation magnitude. By supposing ϵ is small enough, we can have:

$$\mathcal{L}(\theta + \epsilon) - \mathcal{L}(\theta) \approx \nabla \mathcal{L}(\theta)\epsilon. \tag{3.7}$$

Further, we hope to obtain the largest loss change to find the optimal weight perturbation ϵ^*, which is be computed as:

$$\epsilon^* := \arg \max_{\|\epsilon\|_2 \leq \rho} \nabla \mathcal{L}(\theta)\epsilon. \tag{3.8}$$

By following dual norm problem, the optimal ϵ^* can be solved as $\rho \text{sign}(\nabla \mathcal{L}(\theta))$ [137], which is essentially the ∞-norm of the gradient $\nabla \mathcal{L}$ multiplied with a scale parameter ρ. Hence, common sharpness minimization aims to minimize:

$$\mathbb{E}_{(x,y)\sim Q}\mathcal{R} := \mathcal{L}(\theta + \rho \text{sign}(\nabla \mathcal{L}(\theta; (x, y)))) - \mathcal{L}(\theta; (x, y)). \tag{3.9}$$

The intuition is that the perturbation along the gradient norm direction increases the loss value significantly. When training on corrupted distributions, the scarce noisy data scatter sparsely in the high-dimensional space. As a consequence, the neighbor of each datum could not be sufficiently explored, thus producing a sharp loss curve. During testing, the unseen noisy data is likely to fall on an unexplored point with a large loss, further causing inaccurate model predictions.

Therefore, instead of directly applying sharpness minimization on the whole dataset, which leads to poor generalization performance [139] (as demonstrated in Section 3.5.1), we focus on sharpness minimization over the worst-case distribution Q. By conducting the worst-case sharpness minimization, we can enhance the flatness of our classifier. Consequently, when predicting unknown data during the test phase, a flat loss landscape is more likely to produce a low loss than a sharp one; hence, our SharpDRO can generalize better than other DRO methods. However, the robust performance largely depends on the worst-case distribution Q, so next, we explain our worst-case data selection.

3.3.2 Worst-Case Data Selection

Generally, the worst-case data selection focuses on finding the most uncertain data distribution Q from the uncertainty set $\mathcal{Q}$, which is an f-divergence ball from the training distribution P [162, 163, 164]. Most works assume each distribution is distinguishable from the other. However, when the distribution index is not available, it would be very hard to select worst-case data. In this section, we investigate two situations: distribution-aware robust generalization and distribution-agnostic robust generalization.

3.3.2.1 Distribution-Aware Robust Generalization

When given annotations to denote different severities of corruptions, the image data x is paired with class label y and distribution index s. Then, the worst-case distribution Q can be found by identifying the sub-distribution $P_s \in P$ that yields the largest loss. Hence, we can optimize through:

$$\min_{\theta}\Big\{ \max_{\substack{\omega_s; \\ Q=\{\omega_s P_s\}_{s=1}^{S}}} \Big\{ \sum_{(x_i,y_i)\in Q} [\mathcal{L}(\theta, \omega_s; (x_i, y_i))] \Big\} + \sum_{(x_i,y_i)\in Q} [\mathcal{R}(\theta, \omega_s; (x_i, y_i))] \Big\}, \quad (3.10)$$

where ω_s belongs to a $(S-1)$-dimensional probability simplex. The first term simply recovers the learning target of GroupDRO [125, 164] and helps find the worst-case distribution Q. Then, by emphasizing the selected P_s, the second sharpness minimization term can act as a sharpness regularizer. As a result, SharpDRO can learn a flatter loss surface on the worst-case data, thus generalizes better compared to GroupDRO, as discussed in Section 3.4.

3.3.2.2 Distribution-Agnostic Robust Generalization

Due to the annotations being extremely expensive in the real world, a practical challenge is how to learn a robust model without a distribution index. Unlike JTT [140], which trains the model through two stages, we aim to solve this problem more efficiently by detecting the worst-case data during network training. As the corrupted data essentially lie out-of-distribution, we are motivated to conduct OOD detection [165, 166, 35, 36, 56] to find the worst-case data.

Particularly, we re-utilize the weight perturbation ϵ^* to compute an OOD score:

$$\omega_i = \max f(\theta; (x_i)) - \max f(\theta + \epsilon^*; (x_i)), \tag{3.11}$$

where $f(\cdot)$ stands for the c-dimensional label prediction in the label space, whose maximum value is considered as prediction confidence. Intuitively, as the model is much more robust to the clean distribution than the corrupted distribution, the prediction of clean data usually exhibits more stability than scarce noisy data when facing perturbations. Hence, if an example comes from a rarely explored distribution, its prediction certainty would deviate significantly from the original value, thus producing a large OOD score, as shown in Section 3.5.1. Note that the major difference is that we target generalization on worst-case data, but OOD detection aims to exclude OOD data.

To this end, we can construct our worst-case dataset as:

$$Q := \left\{ \sum_{i=1}^{M} \bar{\omega}_i \cdot (x_i, y_i) : \bar{\omega}_i = \frac{\omega_i}{\frac{1}{M}\sum_{i=1}^{M} \omega_i} \right\}, \tag{3.12}$$

where normalization on ω_i is performed simultaneously. Then, the learning target of the distribution-agnostic setting becomes:

$$\min_{\theta} \Big\{ \max_{\bar{\omega}_i} \Big\{ \sum_{(x_i,y_i)\in Q} [\mathcal{L}(\theta, \bar{\omega}_i; (x, y))] \Big\} + \sum_{(x_i,y_i)\in Q} [\mathcal{R}(\theta, \bar{\omega}_i; (x, y))] \Big\}, \tag{3.13}$$

Therefore, the worst-case data can be selected by focusing on the examples with large OOD scores. In this way, our sharpDRO can be successfully deployed into the distribution-agnostic setting to ensure robust generalization, whose effectiveness is demonstrated by quantitative and qualitative results in Sections 3.4.3 and 3.5.1. Next, we give details about implementing SharpDRO.

3.3.3 Optimization for SharpDRO

In both distribution-aware and distribution-agnostic scenarios, the worst-case data distribution is identified using the distribution weighting parameter ω_s and OOD score ω, respectively. Intuitively, their effect is similar: finding the worst data distribution that yields

Algorithm 2 Optimization process of SharpDRO

Training set $\mathcal{D}^{\text{train}} = \{x_i, y_i\}_{i=1}^{M}$ containing Poisson distributed noisy corruptions;
Model parameter $\theta \leftarrow \theta_0$;
Weighting parameter $\omega \leftarrow \omega_0$;
Learning rate: η_θ, η_ω.
for $t = 0, 1, \ldots, T-1$ **do**
  if Distribution-aware **then**
    ▷ Loss maximization via optimizing ω_{t+1}
    $\omega_{t+1} := \arg\max_\omega \left\{\mathbb{E}_{(x,y)\sim\omega P_s}\left[\mathcal{L}(\theta_t, \omega; (x, y))\right]\right\}$;
  else if Distribution-agnostic **then**
    ▷ OOD detection for computing ω_{t+1}
    Update ω_{t+1} via Equation (3.11);
  end if
  ▷ Optimize variable θ
  $\theta_{t+1} = \arg\min_\theta \left\{\mathbb{E}_{(x,y)\sim\omega_t P}[\mathcal{L}(\theta, \omega_t) + \mathcal{R}(\theta, \omega_t)]\right\}$
end for

the maximum loss. Therefore, without loss of generality, we consider the maximization in Equations (3.10) and (3.13) as the optimization on the same weighting parameter ω and the samples (x, y) can be considered i.i.d. from $\mathcal{Q}$ weighted by ω to compute the loss value $\mathcal{L}$. Moreover, the sharpness regularization can be reformulated in the same way as Equation (3.6) by including ω: $\mathcal{R}(\theta, \omega; (x, y)) = \max_{\|\epsilon\|_2 \leq \rho}\{\mathcal{L}(\theta+\epsilon, \omega; (x, y)) - \mathcal{L}(\theta, \omega; (x, y))\}$. Therefore, our general learning objective can be formulated as a bi-level optimization:

$$\min_\theta \mathbb{E}_{(x,y)\sim Q}\left[\mathcal{L}(\theta, \omega^*; (x, y))\right] + \mathcal{R}(\theta, \omega^*; (x, y)) \tag{3.14}$$

$$\text{s. t. } \omega^* = \arg\max_\omega \mathbb{E}_{(x,y)\sim Q}\left[\mathcal{L}(\theta, \omega; (x, y))\right]. \tag{3.15}$$

The optimization process is shown in Algorithm 2. Specifically, we first update the weighting parameter ω based on the empirical risk term $\mathcal{L}$ using stochastic gradient ascent. Then, by leveraging the updated ω, we optimize the general objective, which contains both risk minimization of $\mathcal{L}$ and worst-case sharpness minimization of $\mathcal{R}$. We iterate these processes until convergence, hoping to minimize the risk on the target loss function $\mathcal{L}$ with the worst-case data distribution.

Convergence Analysis: First we give some brief notations:

$$\mathbb{L}(\theta, \omega) := \mathbb{E}_{(x,y)\sim Q}\mathcal{L}(\theta, \omega; (x, y)). \tag{3.16}$$

TABLE 3.1. Quantitative comparisons on distribution-aware robust generalization setting. Averaged accuracy (%) with standard deviations is computed over three independent trials.

Dataset	Type	Method	Corruption Severity					
			0	1	2	3	4	5
CIFAR10	Gaussian	ERM	90.9 ± 0.02	89.2 ± 0.02	86.4 ± 0.03	85.9 ± 0.01	83.5 ± 0.01	78.8 ± 0.01
		IRM	91.8 ± 0.01	90.3 ± 0.01	89.5 ± 0.01	86.7 ± 0.02	81.8 ± 0.02	80.0 ± 0.02
		REx	91.3 ± 0.03	89.5 ± 0.02	88.1 ± 0.02	86.7 ± 0.02	83.3 ± 0.01	80.5 ± 0.02
		GroupDRO	90.2 ± 0.03	89.1 ± 0.02	88.4 ± 0.04	84.3 ± 0.01	83.0 ± 0.02	78.2 ± 0.02
		SharpDRO	$\mathbf{93.1 \pm 0.05}$	$\mathbf{92.2 \pm 0.03}$	$\mathbf{91.4 \pm 0.03}$	$\mathbf{89.2 \pm 0.04}$	$\mathbf{87.1 \pm 0.03}$	$\mathbf{84.5 \pm 0.03}$
	Shot	ERM	$\mathbf{92.5 \pm 0.02}$	91.1 ± 0.02	89.9 ± 0.01	85.6 ± 0.03	85.7 ± 0.01	78.8 ± 0.01
		IRM	90.4 ± 0.01	90.3 ± 0.02	89.4 ± 0.02	86.3 ± 0.01	84.3 ± 0.02	79.1 ± 0.02
		REx	91.1 ± 0.02	90.6 ± 0.02	90.2 ± 0.03	86.8 ± 0.02	84.7 ± 0.02	80.5 ± 0.01
		GroupDRO	92.2 ± 0.01	91.4 ± 0.01	89.4 ± 0.02	84.0 ± 0.01	84.7 ± 0.02	78.3 ± 0.01
		SharpDRO	91.7 ± 0.04	$\mathbf{93.3 \pm 0.04}$	$\mathbf{92.5 \pm 0.05}$	$\mathbf{91.0 \pm 0.02}$	$\mathbf{88.2 \pm 0.03}$	$\mathbf{83.9 \pm 0.04}$
CIFAR100	Gaussian	ERM	68.2 ± 0.01	64.8 ± 0.01	60.6 ± 0.01	56.9 ± 0.01	53.9 ± 0.01	50.2 ± 0.03
		IRM	64.7 ± 0.02	64.7 ± 0.01	62.2 ± 0.01	54.5 ± 0.02	53.4 ± 0.03	50.4 ± 0.01
		REx	68.0 ± 0.03	65.1 ± 0.03	61.8 ± 0.01	56.8 ± 0.01	53.2 ± 0.01	51.5 ± 0.01
		GroupDRO	66.1 ± 0.01	61.7 ± 0.02	59.3 ± 0.03	53.6 ± 0.01	54.0 ± 0.02	50.6 ± 0.02
		SharpDRO	$\mathbf{72.3 \pm 0.03}$	$\mathbf{72.2 \pm 0.03}$	$\mathbf{71.5 \pm 0.02}$	$\mathbf{62.4 \pm 0.03}$	$\mathbf{59.7 \pm 0.03}$	$\mathbf{55.7 \pm 0.06}$
	Shot	ERM	67.6 ± 0.03	65.1 ± 0.01	62.9 ± 0.01	56.0 ± 0.01	55.1 ± 0.01	47.3 ± 0.01
		IRM	67.5 ± 0.02	65.7 ± 0.01	62.7 ± 0.01	59.5 ± 0.01	55.8 ± 0.01	48.3 ± 0.01
		REx	65.7 ± 0.01	63.8 ± 0.02	61.9 ± 0.01	59.3 ± 0.03	53.8 ± 0.01	48.1 ± 0.01
		GroupDRO	67.0 ± 0.02	65.8 ± 0.01	63.1 ± 0.01	58.9 ± 0.01	57.5 ± 0.01	49.3 ± 0.01
		SharpDRO	$\mathbf{71.3 \pm 0.02}$	$\mathbf{70.5 \pm 0.03}$	$\mathbf{68.3 \pm 0.03}$	$\mathbf{63.6 \pm 0.04}$	$\mathbf{60.5 \pm 0.04}$	$\mathbf{53.4 \pm 0.03}$
ImageNet30	Gaussian	ERM	87.5 ± 0.01	84.6 ± 0.01	81.9 ± 0.01	76.5 ± 0.01	71.2 ± 0.01	65.3 ± 0.01
		IRM	86.6 ± 0.01	84.4 ± 0.03	80.6 ± 0.01	75.2 ± 0.01	70.7 ± 0.03	64.8 ± 0.01
		REx	86.3 ± 0.01	83.8 ± 0.03	81.1 ± 0.02	75.6 ± 0.02	71.5 ± 0.01	66.1 ± 0.03
		GroupDRO	85.1 ± 0.02	84.2 ± 0.01	81.2 ± 0.03	76.3 ± 0.03	72.0 ± 0.02	66.3 ± 0.01
		SharpDRO	$\mathbf{89.6 \pm 0.03}$	$\mathbf{88.3 \pm 0.02}$	$\mathbf{85.5 \pm 0.03}$	$\mathbf{82.8 \pm 0.04}$	$\mathbf{77.5 \pm 0.03}$	$\mathbf{70.2 \pm 0.05}$
	Shot	ERM	86.9 ± 0.01	84.8 ± 0.01	83.6 ± 0.01	79.7 ± 0.01	75.4 ± 0.01	64.6 ± 0.01
		IRM	86.8 ± 0.01	85.1 ± 0.03	81.5 ± 0.01	73.5 ± 0.02	68.5 ± 0.03	62.5 ± 0.03
		REx	83.8 ± 0.01	86.3 ± 0.03	82.5 ± 0.02	73.9 ± 0.01	70.6 ± 0.03	64.0 ± 0.02
		GroupDRO	86.7 ± 0.01	85.6 ± 0.03	84.5 ± 0.01	80.7 ± 0.01	76.2 ± 0.04	65.4 ± 0.01
		SharpDRO	$\mathbf{88.3 \pm 0.02}$	$\mathbf{88.1 \pm 0.03}$	$\mathbf{86.4 \pm 0.02}$	$\mathbf{85.3 \pm 0.04}$	$\mathbf{78.2 \pm 0.04}$	$\mathbf{68.2 \pm 0.03}$

The worst-case data distribution, which has the maximum loss, is denoted by $\omega^*(\theta) := \arg\max_\omega \mathbb{L}(\theta, \omega)$. We can obtain the convergence to a stationary point of

$$\mathbb{L}^*(\theta) := \max_{\omega} \mathbb{L}(\theta, \omega) = \mathbb{L}(\theta, \omega^*(\theta)) \tag{3.17}$$

by averaged gradient $\frac{1}{T}\sum_{t=0}^{T-1} \mathbb{E}\|\nabla \mathbb{L}^*(\theta_t)\|^2$.

THEOREM 1 (**Informal**). *Assuming the loss function $\mathbb{L}$ is l-Lipschitz smooth, satisfies μ-Polyak-Łojasiewicz (PL) condition on the second variable ω, and has unbiased estimation about the gradient as well as σ^2 bounded variance, we can get the convergence rate during T iterations:*

$$\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}\|\nabla\mathbb{L}^*(\theta_t)\|^2 \leq 320\sqrt{\frac{3\kappa^4 l(\mathbb{E}[\mathbb{L}^*(\theta_0)]-\min_\theta \mathbb{E}[\mathbb{L}^*(\theta)])\sigma^2}{11MT}} = \mathcal{O}\left(\frac{\kappa^2}{\sqrt{MT}}\right),$$

where the conditional number $\kappa = l/\mu$ and M means the sample batch(here we can choose $M = 1$)[1].

3.4 Experiment

In experiments, we first give details about our experimental setup. Then, we conduct quantitative experiments to compare to proposed SharpDRO with the most popular baseline methods by considering both distribution-aware and distribution-agnostic settings, which shows the capability of SharpDRO to tackle the most challenging distributions. Finally, we conduct qualitative analyses to validate the effectiveness of SharpDRO in robust generalization.

3.4.1 Practical Implementation

Our SharpDRO requires two backward phases, so the time complexity is twice as much as plain training, for efficient sharpness computation, please refer to [167, 168, 169, 170, 171]. In the first step, we record the label prediction p of each data during inference and simultaneously compute the loss $\mathcal{L}$. Additionally, in the first backward pass, we store the computed gradient $\nabla\mathcal{L}(\theta)$. Further, by adding ϵ^*, we use the perturbed model to compute the second label prediction $\hat{p}$, which is further leveraged to compute the sharpness regularization $\mathcal{R}$. Moreover, in the distribution-agnostic setting, the predictions p and $\hat{p}$ from two forward steps are used to compute the OOD score ω_i. Then, we add the recorded gradient $\nabla\mathcal{L}(\theta)$

[1]The resulting bound here means our SharpDRO can converge to the ϵ-stationary point in $\frac{1}{\epsilon^2}$ iterations.

TABLE 3.2. Quantitative comparisons on distribution-agnostic robust generalization setting. Averaged accuracy (%) with standard deviations is computed over three independent trials.

Dataset	Type	Method	Corruption Severity					
			0	1	2	3	4	5
CIFAR10	Gaussian	JTT	89.9 ± 0.02	88.8 ± 0.02	86.5 ± 0.02	86.1 ± 0.02	83.4 ± 0.03	79.8 ± 0.02
		EIIL	88.6 ± 0.02	87.5 ± 0.03	86.3 ± 0.03	85.4 ± 0.02	83.2 ± 0.03	78.8 ± 0.01
		SharpDRO	$\mathbf{91.2 \pm 0.02}$	$\mathbf{91.3 \pm 0.03}$	$\mathbf{88.9 \pm 0.04}$	$\mathbf{87.3 \pm 0.02}$	$\mathbf{85.6 \pm 0.04}$	$\mathbf{83.1 \pm 0.02}$
	Shot	JTT	91.3 ± 0.02	90.5 ± 0.03	89.3 ± 0.01	86.5 ± 0.02	83.1 ± 0.02	79.8 ± 0.02
		EIIL	90.3 ± 0.03	90.1 ± 0.02	88.3 ± 0.01	86.2 ± 0.02	82.3 ± 0.03	78.5 ± 0.02
		SharpDRO	$\mathbf{91.9 \pm 0.02}$	$\mathbf{91.1 \pm 0.02}$	$\mathbf{90.2 \pm 0.04}$	$\mathbf{88.6 \pm 0.04}$	$\mathbf{86.5 \pm 0.05}$	$\mathbf{83.3 \pm 0.04}$
CIFAR100	Gaussian	JTT	68.0 ± 0.02	65.3 ± 0.02	61.3 ± 0.01	56.3 ± 0.01	54.2 ± 0.03	51.2 ± 0.02
		EIIL	67.2 ± 0.01	66.2 ± 0.02	61.0 ± 0.02	55.8 ± 0.02	54.6 ± 0.03	52.1 ± 0.02
		SharpDRO	$\mathbf{70.3 \pm 0.03}$	$\mathbf{68.8 \pm 0.03}$	$\mathbf{65.2 \pm 0.03}$	$\mathbf{60.3 \pm 0.02}$	$\mathbf{57.4 \pm 0.03}$	$\mathbf{55.3 \pm 0.03}$
	Shot	JTT	66.3 ± 0.02	65.3 ± 0.03	63.4 ± 0.02	56.6 ± 0.04	55.5 ± 0.04	48.6 ± 0.04
		EIIL	66.5 ± 0.02	65.3 ± 0.03	62.8 ± 0.04	57.5 ± 0.02	56.5 ± 0.01	49.5 ± 0.01
		SharpDRO	$\mathbf{68.9 \pm 0.02}$	$\mathbf{66.2 \pm 0.03}$	$\mathbf{64.9 \pm 0.03}$	$\mathbf{59.8 \pm 0.02}$	$\mathbf{56.5 \pm 0.03}$	$\mathbf{51.0 \pm 0.02}$
ImageNet30	Gaussian	JTT	87.3 ± 0.02	84.5 ± 0.02	82.3 ± 0.04	75.6 ± 0.01	72.1 ± 0.04	66.5 ± 0.02
		EIIL	$\mathbf{88.2 \pm 0.02}$	85.2 ± 0.03	81.3 ± 0.02	74.5 ± 0.02	71.5 ± 0.02	65.0 ± 0.04
		SharpDRO	87.5 ± 0.03	$\mathbf{86.6 \pm 0.03}$	$\mathbf{85.3 \pm 0.03}$	$\mathbf{79.3 \pm 0.04}$	$\mathbf{75.3 \pm 0.02}$	$\mathbf{70.0 \pm 0.02}$
	Shot	JTT	86.5 ± 0.02	85.4 ± 0.03	82.6 ± 0.04	79.6 ± 0.04	77.2 ± 0.04	65.0 ± 0.01
		EIIL	85.5 ± 0.01	86.3 ± 0.04	81.6 ± 0.02	80.2 ± 0.03	75.3 ± 0.02	64.4 ± 0.03
		SharpDRO	$\mathbf{87.1 \pm 0.02}$	$\mathbf{87.1 \pm 0.03}$	$\mathbf{84.8 \pm 0.04}$	$\mathbf{83.0 \pm 0.02}$	$\mathbf{76.5 \pm 0.03}$	$\mathbf{69.2 \pm 0.04}$

back to the model parameter and conduct sharpness minimization over the selected worst-case data. In this way, our SharpDRO can be correctly performed.

3.4.2 Experimental Setup

For distribution-aware situation, we choose GroupDRO [125], IRM [69], REx [144], and ERM for comparisons. As for a distribution-agnostic situation, we pick JTT [140] and Environment Inference for Invariant Learning (EIIL) [143] for baseline methods[2]. For each problem setting, we construct corruption using CIFAR10/100 [107] and ImageNet30 [104] datasets. Specifically, we follow [114] to perturb the image data with severity levels varying from 1 to 5 by using two types of corruption: "Gaussian Noise" and "Shot Noise".

[2]Note that we do not include the sharpness minimization method SAM [137] in this problem setting because its OOD generalization performance is worse than ERM. However, we conduct a detailed analysis between SharpDRO and SAM in Section 3.5.1

TABLE 3.3. Quantitative comparisons on distribution-aware robust generalization setting on Snow corruption. Averaged accuracy (%) with standard deviations are computed over three independent trails.

Dataset	Type	Method	Corruption Severity 0	1	2	3	4	5
CIFAR10	Snow	ERM	90.8 ± 0.01	90.1 ± 0.02	88.1 ± 0.02	88.1 ± 0.02	85.7 ± 0.02	82.6 ± 0.01
		IRM	91.1 ± 0.02	90.7 ± 0.01	89.7 ± 0.02	88.0 ± 0.03	84.6 ± 0.02	83.2 ± 0.03
		REx	91.8 ± 0.02	$\mathbf{91.9 \pm 0.01}$	88.4 ± 0.01	88.3 ± 0.01	88.6 ± 0.01	83.0 ± 0.02
		GroupDRO	91.5 ± 0.02	91.0 ± 0.01	88.7 ± 0.02	88.6 ± 0.02	85.2 ± 0.03	83.5 ± 0.02
		SharpDRO	$\mathbf{93.1 \pm 0.01}$	91.8 ± 0.01	$\mathbf{90.5 \pm 0.02}$	$\mathbf{90.8 \pm 0.02}$	$\mathbf{87.9 \pm 0.01}$	$\mathbf{84.3 \pm 0.02}$
CIFAR10	Snow	ERM	67.7 ± 0.01	68.1 ± 0.01	64.7 ± 0.01	63.1 ± 0.01	60.5 ± 0.02	57.3 ± 0.01
		IRM	69.3 ± 0.01	67.5 ± 0.02	64.9 ± 0.02	61.0 ± 0.01	58.2 ± 0.01	55.1 ± 0.01
		REx	66.4 ± 0.01	65.9 ± 0.01	62.4 ± 0.01	61.2 ± 0.02	57.5 ± 0.03	56.0 ± 0.02
		GroupDRO	68.0 ± 0.02	68.2 ± 0.01	65.1 ± 0.01	60.9 ± 0.03	59.8 ± 0.01	58.1 ± 0.02
		SharpDRO	$\mathbf{71.5 \pm 0.01}$	$\mathbf{70.8 \pm 0.03}$	$\mathbf{67.5 \pm 0.02}$	$\mathbf{65.5 \pm 0.01}$	$\mathbf{62.3 \pm 0.01}$	$\mathbf{59.2 \pm 0.03}$
ImageNet30	Snow	ERM	86.7 ± 0.03	85.2 ± 0.01	83.4 ± 0.01	81.1 ± 0.01	75.3 ± 0.01	75.6 ± 0.01
		IRM	85.6 ± 0.01	84.0 ± 0.02	82.1 ± 0.03	79.7 ± 0.01	75.0 ± 0.01	75.6 ± 0.01
		REx	85.4 ± 0.01	84.6 ± 0.02	82.7 ± 0.02	80.5 ± 0.03	75.7 ± 0.03	75.9 ± 0.03
		GroupDRO	86.7 ± 0.01	85.5 ± 0.03	83.4 ± 0.01	81.2 ± 0.02	76.3 ± 0.01	76.6 ± 0.01
		SharpDRO	$\mathbf{88.2 \pm 0.02}$	$\mathbf{88.2 \pm 0.01}$	$\mathbf{85.4 \pm 0.02}$	$\mathbf{81.9 \pm 0.01}$	$\mathbf{79.8 \pm 0.03}$	$\mathbf{79.5 \pm 0.02}$

Moreover, the clean data are considered as having a corruption severity of 0. For each corrupted distribution, we sample them with different probabilities by following the Poisson distribution $P(s; \lambda = 1)$, i.e., for s varies from 0 to 5, the sample probabilities are $\{0.367, 0.367, 0.184, 0.061, 0.015, 0.003\}$, respectively. Then, we test the robust performance on each data distribution. For hyper-parameter ρ, we follow [137] by setting it to 0.05 to control the magnitude of ϵ^*. For each experiment, we conduct three independent trials and report the average test accuracy with standard deviations.

3.4.3 Quantitative Comparisons

In this part, we focus on three questions: (1) Can SharpDRO perform well in two situations of robust generalization? (2) Does SharpDRO generalize well on the most severely corrupted distributions? and (3) Is SharpDRO able to tackle different types of corruption? To answer these questions, we conduct experiments on both settings by testing on different corruption types and severity levels.

TABLE 3.4. Quantitative comparisons on distribution-agnostic robust generalization setting on snow corruption. Averaged accuracy (%) with standard deviations are computed over three independent trails.

Dataset	Type	Method	Corruption Severity 0	1	2	3	4	5
CIFAR10	Snow	JTT	88.6 ± 0.02	87.8 ± 0.03	86.5 ± 0.02	87.2 ± 0.02	84.2 ± 0.02	83.2 ± 0.03
		EIIL	88.3 ± 0.02	87.8 ± 0.01	85.6 ± 0.02	87.3 ± 0.03	85.2 ± 0.04	82.3 ± 0.01
		SharpDRO	$\mathbf{91.6 \pm 0.01}$	$\mathbf{91.1 \pm 0.02}$	$\mathbf{90.8 \pm 0.01}$	$\mathbf{89.7 \pm 0.02}$	$\mathbf{86.2 \pm 0.01}$	$\mathbf{83.8 \pm 0.02}$
CIFAR10	Snow	JTT	67.5 ± 0.01	68.1 ± 0.02	65.3 ± 0.02	64.3 ± 0.02	60.2 ± 0.02	57.8 ± 0.02
		EIIL	68.2 ± 0.03	69.1 ± 0.03	65.2 ± 0.02	64.0 ± 0.02	61.0 ± 0.04	57.5 ± 0.04
		SharpDRO	$\mathbf{70.6 \pm 0.02}$	$\mathbf{69.9 \pm 0.03}$	$\mathbf{66.7 \pm 0.03}$	$\mathbf{64.4 \pm 0.02}$	$\mathbf{61.9 \pm 0.03}$	$\mathbf{60.7 \pm 0.03}$
ImageNet30	Snow	JTT	86.0 ± 0.04	85.8 ± 0.02	82.3 ± 0.03	80.4 ± 0.02	74.6 ± 0.02	73.5 ± 0.02
		EIIL	87.5 ± 0.01	85.4 ± 0.02	83.5 ± 0.04	$\mathbf{81.6 \pm 0.01}$	76.3 ± 0.01	75.8 ± 0.02
		SharpDRO	$\mathbf{87.5 \pm 0.03}$	$\mathbf{86.7 \pm 0.02}$	$\mathbf{85.4 \pm 0.02}$	81.5 ± 0.03	$\mathbf{78.9 \pm 0.02}$	$\mathbf{78.5 \pm 0.03}$

Distribution-Aware Robust Generalization As shown in Table 3.1, we can see that SharpDRO surpasses other methods with larger performance gains as the corruption severity increases. Especially in the CIFAR10 dataset on "Gaussian Noise" corruption, the improvement margin between SharpDRO and the second-best method is 2.2% with severity of 0, which is further increased to about $\mathbf{5.7}\%$ with severity of 5, which indicates the capability of SharpDRO on generalization against severe corruptions. Moreover, SharpDRO frequently outperforms other methods on all scenarios, which manifests the robustness of SharpDRO against various corruption types.

Distribution-Agnostic Robust Generalization In Table 3.2, we can see a similar phenomenon as in Table 3.1 that the more severe corruptions are applied, the larger performance gains SharpDRO achieves. Especially, in the ImageNet30 dataset corrupted by "Shot Noise", SharpDRO shows about 0.6% performance gains upon the second-best method with severity 0, which is further increased to almost $\mathbf{4.2}\%$ with severity 5. Moreover, SharpDRO is general to all three corruption types, as it surpasses other methods in most cases. Therefore, SharpDRO can perfectly generalize even without the distribution annotations.

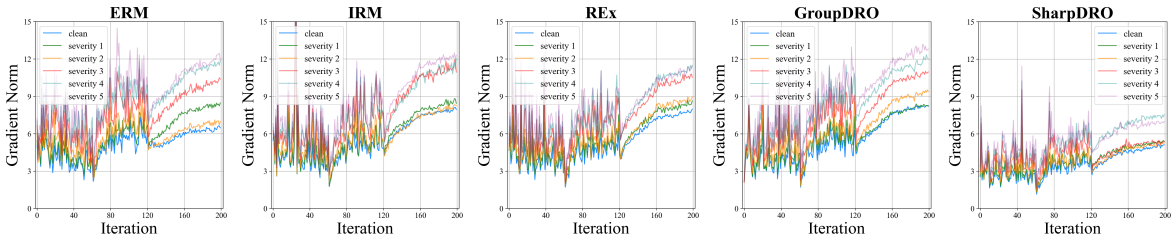


FIGURE 3.4. Gradient norm comparisons between different methods over all corrupted distributions.

3.5 Results on Additional Corruptions

In the main paper, we have provided the results using "Gaussian Noise" corruption and "Shot Noise" corruption. Here, we conduct additional experiments to show the effectiveness of SharpDRO under "Snow" corruption. The results on CIFAR10, CIFAR100, and ImageNet30 datasets in both distribution-aware and distribution-agnostic scenarios are shown in Tables 3.3 and 3.4. We can see that SharpDRO still performs effectively and surpasses other methods with a large margin. Especially, on the ImageNet30 dataset in both problem settings, SharpDRO outperforms the second-best method by about 3%, which indicates the capability of SharpDRO on generalization against different corruptions.

3.5.1 Qualitative Analysis

To investigate the effectiveness of SharpDRO, we first conduct an ablation study to show that the worst-case sharpness minimization is essential for achieving generalization with robustness. Then, we utilize gradient norm, an important criterion to present training stability, to validate that our method is stable for severely corrupted distributions. Then, we analyze the hyper-parameter ρ and OOD score $\bar{w}$ to disclose the effectiveness of sharpness minimization and worst-case data selection. Finally, another second-order method, SAM [137, 172, 173, 174, 175], is investigated to discover the efficiency property of SharpDRO. All analyses are conducted using CIFAR10 with "Gaussian Noise" corruption.

TABLE 3.5. Ablation study. "w/o data selection" denotes training without worst-case data selection, which recovers SAM [137], and "w/o sharp min" indicates training without sharpness minimization, which is the same as GroupDRO [125].

Method	Corruption Severity					
	0	1	2	3	4	5
w/o data selection (SAM)	93.2	90.5	87.6	82.1	80.5	75.4
w/o sharp min (GroupDRO)	90.2	89.1	88.4	84.3	83.0	78.2
SharpDRO	93.1	92.2	91.4	89.2	87.1	84.5

Ablation Study By eliminating the worst-case data selection, we recover the original sharpness minimization method SAM [137]. Then, we remove the sharpness minimization module, which is basically training via GroupDRO. The ablation results are shown in Table 3.5. We can see that deploying SAM on the whole training dataset can achieve improved results on the clean dataset. However, the robust performance on corrupted distributions is even worse than GroupDRO. This could be because sharpness is easy to be dominated by principal distributions, which is misleading for generalization to small distributions. Thus, the sharpness of corrupted data would be sub-optimal. As for GroupDRO, it fails to produce a flat loss surface, hence it cannot generalize as well as the proposed SharpDRO.

Distributional Stability To show our method can be stable even in the most challenging distributions, we show the gradient norm on a validation set including corruption severity from 0 to 5. As shown in Figure 3.4, SharpDRO not only produces the smallest norm value but also can ensure almost equal gradient norm across all corruptions, which indicates that SharpDRO is the most distributionally stable method among all compared methods.

Parameter Analysis To understand how the scale parameter ρ affects our generalization performance, we conduct a sensitivity analysis by changing this value and show the test results of different distributions. In Figure 3.5 (a), we find an interesting discovery that as ρ increases, which indicates the perturbation magnitude ϵ^* enlarges, would enhance the generalization of severely corrupted data but degrade the performance of slightly corrupted data. This might be because the exploration of hard distributions needs to cover wa ide range of neighborhoods to ensure generalization. On the contrary, exploration too far on easy

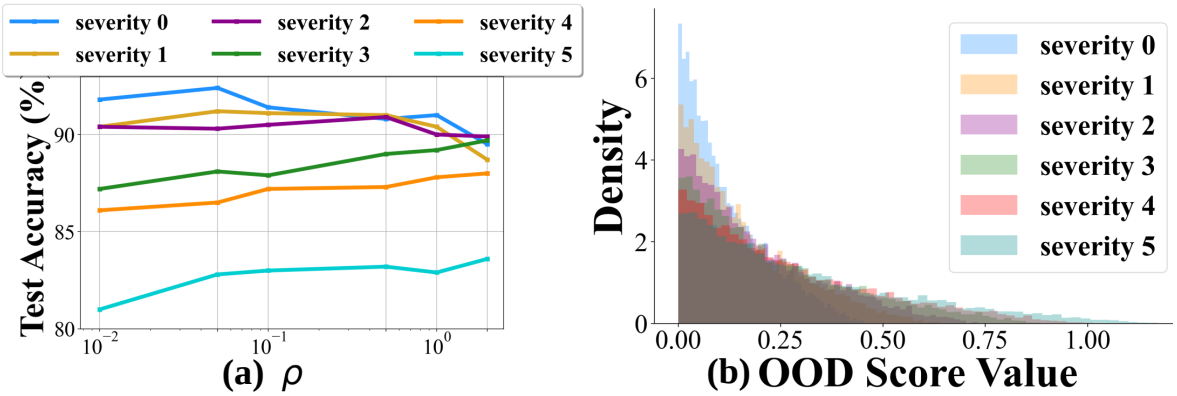


FIGURE 3.5. (a) Sensitivity of ρ whose value is set to $\{0.01, 0.05, 0.1, 0.5, 1, 2\}$. (b) Distribution of the normalized OOD score $\bar{\omega}$ on distribution $s = 0$ to 5.

distributions can reach out-of-distribution, thus causing performance degradation. Therefore, for practitioners who aim to generalize on small and difficult datasets, we might be able to enhance performance by aggressively setting a large perturbation scale.

OOD Score Analysis The OOD score helps to select worst-case data under the distribution-agnostic setting. To show its effectiveness in selecting the noisy data, we plot the value distribution of OOD scores from all corrupted distributions in epoch 30 in Figure 3.5 (b). We can see the tendency that severe corruption has larger OOD scores. Therefore, our OOD score is a valid criterion to select worst-case data. Note that during the training process, the worst-case data would be gradually learned, thus the OOD score can become smaller, which explains why the value distribution of our score is not as separable as OOD detection does.

Training Efficiency Analysis It is clear that the proposed SharpDRO method is a second-order optimization method. Hence, when compared to first-order methods such as GroupDRO and REx, computational cost is the price to pay for achieving improved generalization performance[3]. However, to further explore the advantage of SharpDRO compared to other second-order methods, here we use SAM [137] as a competitor, and show their computational time as well as worst-case accuracy ($s = 5$) in Figure 3.6. We can see that on all three datasets, our SharpDRO requires nearly the same time to train, and significantly outperforms

[3]Note that our method can be deployed with existing efficient sharpness-based methods [169, 167, 168, 171].

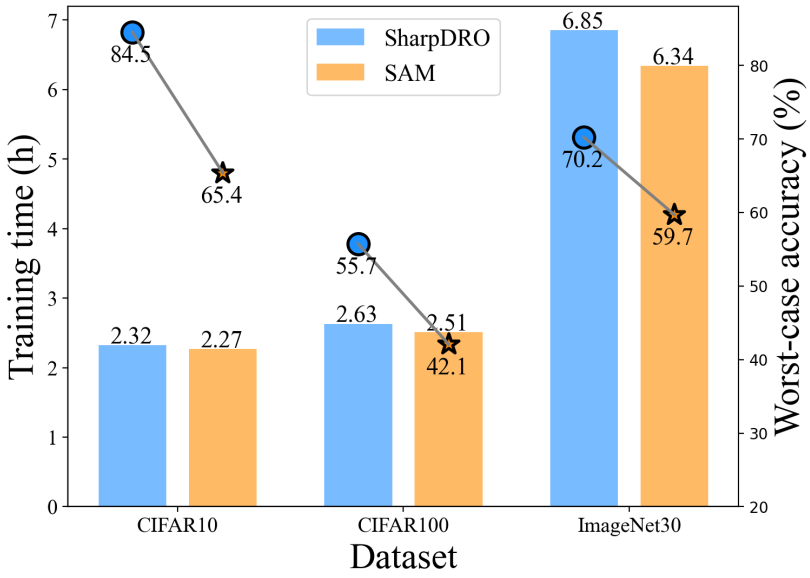


FIGURE 3.6. Efficiency comparison between SharpDRO and SAM.

the worst-case performance of SAM, owing to our efficient worst-case data selection, which is vital for robust generalization against severe corruptions.

3.6 Conclusion

In this Chapter, we propose a SharpDRO approach to enhance the generalization performance of DRO methods. Specifically, we focus on minimizing the sharpness of worst-case data to learn flat loss surfaces. As a result, SharpDRO is more robust to severe corruptions compared to other methods. Moreover, we apply SharpDRO to distribution-aware and distribution-agnostic settings and propose an OOD detection process to select the worst-case data when the distribution index is not known. Extensive quantitative and qualitative experiments have been conducted to show that SharpDRO can deal with the most challenging corrupted distributions and achieve improved generalization results compared to well-known baseline methods.

CHAPTER 4

Exploring Variant Parameters for Invariant Learning

Out-of-Distribution (OOD) Generalization aims to learn robust models that generalize well to various environments without fitting to distribution-specific features. Recent studies based on Lottery Ticket Hypothesis (LTH) address this problem by minimizing the learning target to find some of the parameters that are critical to the task. However, in OOD problems, such solutions are suboptimal as the learning task contains severe distribution noises, which can mislead the optimization process. Therefore, apart from finding the task-related parameters (i.e., invariant parameters), we propose Exploring Variant parameters for Invariant Learning (EVIL) which also leverages the distribution knowledge to find the parameters that are sensitive to distribution shift (i.e., variant parameters). Once the variant parameters are left out of invariant learning, a robust subnetwork that is resistant to distribution shift can be found. Additionally, the parameters that are relatively stable across distributions can be considered invariant ones to improve invariant learning. By fully exploring both variant and invariant parameters, our EVIL can effectively identify a robust subnetwork to improve OOD generalization. In extensive experiments on integrated testbed: DomainBed, EVIL can effectively and efficiently enhance many popular methods, such as ERM, IRM, SAM, etc.

4.1 Introduction

The strong representation ability of deep neural networks [3, 107, 1] has been one of the vital keys to the success of deep learning over the past decade. However, the realistic deployment of neural networks is often restricted to the IID assumption, where the training data and test data should be distributed independently and identically. When such an assumption is

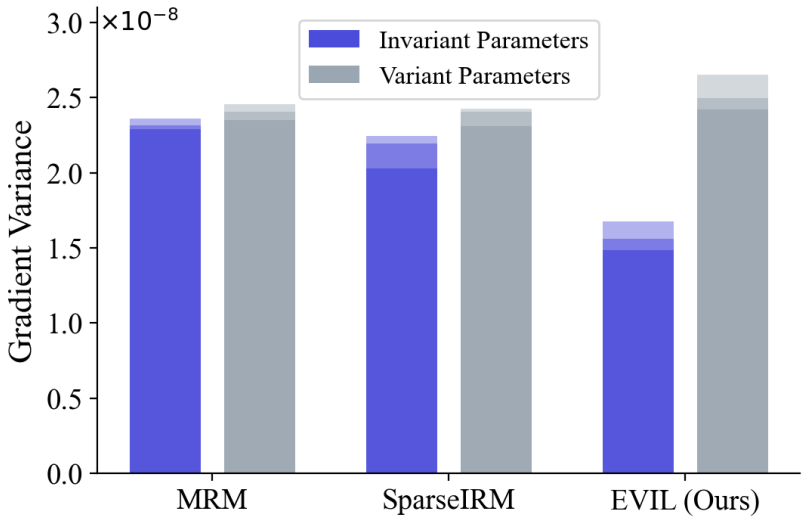


FIGURE 4.1. Comparison of the gradient variance between the learned subnetwork, i.e., invariant parameters, and the pruned parameters, i.e., variant parameters. The gradient variance is computed through $\mathcal{V} = Mean(Var([g_i]_{i=0}^{d}))$ [176], where g_i denotes the i-th gradient among d distributions, and $Var(\cdot)$ and $Mean(\cdot)$ denotes the mathematical variance and mean, respectively. The results are from three independent trials.

violated, a drastic degradation in learning performance is often observed, which seriously hinders the practical application of deep models. Therefore, Out-Of-Distribution (OOD) generalization [177, 178, 152] thrives as a promising direction that aims to enhance model robustness against unknown distribution shifts.

In order to achieve OOD generalization, one mainstream methodology is invariant learning [69, 143, 179, 144], which enforces extracting invariant features to help make consistent predictions among various data distributions (or domains), meanwhile avoiding learning distribution-specific features that are irrelevant to label information. Recent advances based on Lottery Ticket Hypothesis (LTH) [180, 181, 182] show that sparse training optimized by learning task could select some critical parameters as a subnetwork, which are strongly responsible for invariant learning [183, 184, 185]. However, in OOD problems, the sparsification guided by the learning task is problematic because the distribution noise could be erroneously incorporated into the optimization of sparse learning. As a consequence, existing methods fail to identify a robust subnetwork that is stable across different distributions.

Particularly, we follow Rame et al. [176] by using gradient variance to indicate model sensitivity to distribution shift. Then, we compare the gradient variance between the subnetwork and the pruned parameters learned by different methods, as shown in Figure 4.1. We can see that the subnetwork learned by existing methods (MRM [184] and SparseIRM [185]) are almost as sensitive as the pruned parameters, which means that invariant information could not be fully captured.

To overcome this problem, we propose a novel sparse training framework by Exploring Variant parameters for Invariant Learning (EVIL), which presents the life philosophy: facing the evil side to achieve higher virtue. Specifically, by following common assumptions that input data can be decomposed into invariant features and spurious features [93, 94, 184, 26][1], we can divide the network parameters into two types: *invariant parameters* that are strongly related to invariant features, and *variant parameters* that can mistakenly produce spurious features. Intuitively, the invariant parameters and variant parameters are mutually exclusive of each other, as they are either helpful or harmful to our learning task. In order to correctly identify an ideal subnetwork for OOD generalization, our EVIL method not only selects invariant parameters based on the learning task but also explores variant parameters via discriminating each distribution, i.e., classifying the data based on the distribution information. In this way, the connection between variant parameters and spurious features can be successfully established. By finding those variant parameters that are strongly activated by the distribution information, we can be sure that they should not be identified as invariant ones, which provides an alternative and effective way to improve invariant learning.

Furthermore, to dynamically improve our identification of invariant parameters during the course of network training, we propose to revisit some parameters that hardly vary when facing distribution shifts. Concretely, starting from an initialized partition of invariant parameters and variant ones, we select some variant parameters that show low response to the distribution information, as they might be critical for learning distribution-invariant features. Hence, such parameters are recollected as invariant ones to learn from label information. On the other hand, some invariant parameters that are insensitive to our learning task shall be

[1]Though some works investigate more complex situations where there are multiple factors causing the data generation process [186, 179, 187], our assumption is more common in OOD generalization.

rejected from sparse training, as they hardly contribute to invariant learning. Through this dynamic process, we are able to identify a robust subnetwork that is stable across different distributions. As shown in Figure 4.1, the invariant parameters learned by our EVIL method show much smaller gradient variance than the rest variant parameters, which manifests its effectiveness for capturing the invariant information.

By applying our EVIL framework to many existing OOD generalization methods, we conduct extensive empirical comparison and analysis to show that EVIL brings promising improvement with little computation cost. Specifically, when combined with simple ERM, our method achieves 2.4% gains on averaged performance from DomainBed. Furthermore, our EVIL framework can surpass existing sparse training methods for invariant learning by a large margin in various sparsity levels.

To sum up, our contributions are threefold:

- We propose a novel sparse training framework for OOD generalization, which can fully explore the variant parameters to capture invariant information.
- An iterative strategy is designed to dynamically improve the identification of robust subnetworks.
- The proposed EVIL framework can be deployed to many popular methods with great effectiveness and efficiency. Moreover, EVIL effectively surpasses existing sparse invariant learning methods.

4.2 Related Work

Invariant learning for OOD generalization seeks to enforce model predictive invariance when facing distribution shifts [69, 42, 43, 188]. Invariant Risk Minimization (IRM) [69] tries to find an optimal classifier for each data distribution such that the spurious information from each domain is left out. Then, Distributionally Robust Optimization (DRO) [164, 189, 125, 190] proposes to tackle the most challenging distribution to improve OOD generalization, which is shown to be effective by using strong regularization penalties. Moreover,

Sharpness-Aware Minimization (SAM) [137] hopes to learn a flat loss landscape via penalizing the sharpness measurement to improve generalization results [139] and robustness to label noise [191, 192, 193]. Further, Risk Extrapolation (REx) [144] finds that only focusing on one of the known distributions might not help generalize to unknown distributions. Instead, REx shows it is beneficial to enforce comparable performance among all training distributions via penalizing risk variance. Additionally, other methods draw insights from causality [60, 61, 153, 194] to disentangle the invariant features from spurious ones [84, 179, 93] so that prediction would not be affected by distribution shift [195, 196].

Nonetheless, existing invariant learning methods suffer from two major drawbacks. Firstly, some of them a computationally expensive. For example, SAM requires second-order computation to manipulate gradient information, and causality-based methods often require training generative models, which is hard to deploy on large-scale datasets. Secondly, as found out by Gulrajani et al. [177], most methods have limited performances which are even worse than Empirical Risk Minimization (ERM)! However, our EVIL can not only avoid redundant optimization on the variant parameters but also fully capture the invariant feature to achieve superior generalization accuracy.

Sparse training for OOD generalization is first brought out by Morcos et al. [183], which aims to discover the generalization ability of sparse networks obtained via common initialization methods. Then, Modular Risk Minimization (MRM) [184] shows that sparse training can possibly improve the OOD generalization performance compared to the original dense network. However, MRM is designed in a static way, which cannot be optimized along with network training, hindering the sparse learning results. To tackle this issue, Sparse Invariant Risk Minimization (SparseIRM) [185] proposes to conduct the sparse training process and IRM simultaneously. As a result, its generalization performance is further improved compared to MRM.

Despite the improvement of existing sparse methods, they are still suboptimal as the sparse training could be affected by the noisy gradient from the learning task. Meanwhile, the

pruned parameters are not properly leveraged, which would cause non-negligible information loss. Fortunately, EVIL can fully explore both variant and invariant parameters dynamically. Thus, it finds an ideal subnetwork minimally influenced by distribution shift.

4.3 A Critical Analysis of Sparse Training with OOD Data

OOD generalization aims to learn an invariant predictor by leveraging multiple distributions of training data such that the generalization performance on unseen test data distributions. Practically, we usually have multiple datasets correspondingly drawn from m distributions (also termed environment), $\mathcal{E} = \{e_1, \dots e_m\}$, where each distribution $e = \{(\mathbf{x}_i, y_i)\}_{i=0}^{n}$ contains n examples $\mathbf{x} \in X \in \mathbb{R}$ with class label $y \in Y \in \mathbb{R}^c$. Therefore, for each example from distribution e, we can assign a distribution index $d \in \mathbb{R}^m$ and denote a data point as $(\mathbf{x}, y, d)$. Moreover, we have a test dataset sampled from unseen distributions $\mathcal{E}_{unseen}$ to evaluate the generalization performance of our invariant learning. Let $f_\theta : \mathbb{R} \to \mathbb{R}^D$ be a parameterized model with parameters $\theta \in \Theta$ which extracts feature $Z \in \mathbb{R}^M$. Our goal is to prune a sparse subnetwork from an overparameterized model so that variant features can be excluded from making the final prediction. Therefore, OOD generalization can be improved.

Data Generation Process. By following the same formulation and problem setting from Zhang et al. [184], we assume the input variable X^e from environment e is generated from latent variables $Z^e = (Z^e_{inv}, Z^e_{var})$. Intuitively, the input X^e indicates the image pixels, while Z^e_{inv} stands for the feature of the object-of-interest that stays invariant across different environments, and Z^e_{var} denotes the spurious feature which is introduced by the change of environments. Then, the data is generated through $X^e = G(Z^e_{inv}, Z^e_{var})$ where $G(\cdot)$ denotes the data generating function. To obtain an OOD-robust model, we hope to extract learning representations Z^e which can recover the invariant feature Z^e_{inv}, meanwhile excluding the variant feature Z^e_{var}. Such a process is modeled through $Z^e = f_\theta(X^e)$, where we hope $Z^e \approx [Z^e_{inv}, \mathbf{0}]$. Hence, based on the extracted feature, we can make predictions through a classification head $\hat{Y}^e = h(Z^e)$ and train our model by minimizing the error between the prediction $\hat{Y}^e$ and ground truth label Y^e.

The Cause of Data Bias. Based on the data generation process, here we explain why different distributions contain biases that hinder the generalization result. We consider a simple example where Z_{inv}^e and Z_{var}^e are multivariate variables with binary elements, i.e., $Z_{inv}^e \in \{-1,1\}^{M_{inv}}$ and $Z_{var}^e \in \{-1,1\}^{M_{var}}$, in which M_{inv} and M_{var} denotes the dimension of invariant feature and spurious feature, respectively. We have class label $Y^e \in \{-1,1\}$ and distribution index $D \in \{-1,1\}$. Since the invariant feature stays constant across environments, we assume each element of Z_{inv}^e is equal to Y^e. On the other hand, we assume each element in Z_{var}^e takes a value equal to Y^e with probability p^e and $-Y^e$ with probability $1-p^e$ [184]. When p^e is large, the spurious feature would be closely correlated with the class label, hence being unlikely to introduce large data biases. Conversely, if p^e is small, Z_{var}^e can easily introduce noisy signals that might flip the prediction.

Additionally, we analyze the domain knowledge to provide an opposite perspective, which is overlooked by previous works [184, 185]. Concretely, the change of distribution index D is the cause of introducing a spurious feature, i.e., $D \rightarrow Z_{var}^e$, as described by many proposed causal structures [84, 197, 198]. Therefore, when given the distribution D, we can find a specific type of spurious feature. Hence, we assume each element of Z_{var}^e is equal to D. On the other hand, we consider Z_{inv}^e takes the value of D with probability q^e and $-D$ with probability $1-q^e$. It has been commonly assumed that the invariant feature and domain knowledge are independent [84], thus the probability q^e could approximately be 0.5.

The Flaw of Common Sparse Training Strategy Existing studies on sparse invariant learning [184, 185] have shown that when pruning an overparameterized model, the OOD generalization performance could be improved substantially. However, we find that the existing pruning strategy, which is based on ERM or objectives only related to labels, could be suboptimal. Specifically, we consider the same data setting described above, $Z_{inv}^e \in \{-1,1\}^{M_{inv}}$ and $Z_{var}^e \in \{-1,1\}^{M_{var}}$. The data generating function G is simplified as an identity map [118, 199], thus $X = (Z_{inv}^e, Z_{var}^e)$. Suppose the classification model f_θ is a linear layer, we have a mask $\mathbf{m}$ randomly initialized with $0-1$ values to prune the parameter θ, and its sparsity ratio is set to $R = \frac{M_{var}}{M}$. Particularly, we denote the selected invariant parameters as $\theta_{inv} = \mathbf{m} \circ \theta$ and the pruned variant parameters as $\theta_{var} = (\mathbf{1} - \mathbf{m}) \circ \theta$ where

$\circ$ is the element-wise production. To ease the calculation, let the parameter values follow a unit norm, i.e., $\theta = \mathbf{1}\frac{1}{\sqrt{M}}$[2].

PROPOSITION 2. *Consider a biased dataset described above, where $Z^e_{inv} \in \{-1, 1\}^{M_{inv}}$ and $Z^e_{var} \in \{-1, 1\}^{M_{var}}$. Let mask $\mathbf{m}$ be randomly initialized to $\{0, 1\}$ values with sparsity ratio $R = \frac{M_{var}}{M}$, and assume Z^e is a multivariate variable with independent elements. For a common sparse training strategy that aims to minimize empirical risk:*

$$Err^e = \frac{1}{2}\mathbb{E}_{(X^e,Y^e)\sim e}\left[1 - Y^e\hat{Y}^e\right], \tag{4.1}$$

we have:

- *The common strategy fails to find invariant parameters, i.e., $\mathbf{m}_{i\in\{0,\dots,M_{inv}\}}$ remains unupdated. When leveraging domain knowledge with regularization:*

$$Err^d = \frac{1}{2}\mathbb{E}_{(X,Y)\sim\mathcal{E}}\left[1 - D\hat{D}\right], \tag{4.2}$$

  the invariant parameters can be effectively selected with probability at least $1 - \frac{q^e}{2}$;
- *On an unknown distribution, the performance of the common strategy is highly sensitive to p^e: $Err^e \leq \mathcal{O}(e^{-(p^e)^4})$, while leveraging domain knowledge achieves a tighter error bound when p^e is small: $Err^e \leq \mathcal{O}(e^{-(p^e)^2})$.*

As we find out, the pruning strategy is not sufficient to find an ideal subnetwork that can exclude spurious features meanwhile extracting invariant features. This is because the invariant parameters do not produce any error. As a result, existing strategies based on connection sensitivity [200], weight value [201], and Fisher information [202] could be suboptimal when dealing with OOD problems because the gradient information is not actually related to invariant parameters, but variant parameters. Based on this intuition, we proposed a simple yet effective strategy that leverages an additional domain knowledge regularization to explore the invariant parameters. Thanks to such a regularization, the invariant parameters can

[2]Note that our assumption is more general than that from Zhang et al. [184], in which only two extreme case are considered: an optimal sparse invariant network only extracts invariant feature and a network completely depending on spurious feature. Our assumption is practical since it is similar to an initial state where all the parameters are initialized with unit-norm.

be selected because they generate gradients when calculating the distribution regularization, thus easy to find. Meanwhile, the variant parameters can still be excluded to avoid learning spurious features. Moreover, based on the error bounds, our method is insensitive to the spurious correlation $1 - p^e$ compared to the common strategy. In a difficult scenario where p^e is small, our method can still be robust to distribution shift.

4.4 Methodology

In this section, we introduce our EVIL framework as shown in Figure 4.2. In the learning flow of EVIL, there are two procedures: Parameter Exploration, in which we propose to not only study invariant parameters but also explore the variant ones; and Invariant Learning, where we train the identified subnetwork to optimize the invariant parameters.

In the following content, we first introduce our EVIL framework, which contains the aforementioned two procedures. Then, we carefully demonstrate the realization of EVIL using an important optimization method: SAM [137], which largely improves OOD generalization.

4.4.1 The Proposed EVIL Framework

In order to get a good initialization, a few steps of pre-training are commonly conducted by minimizing a learning objective $\mathcal{L}(f_\theta(\mathbf{x}))$ [203, 204, 201, 183], where f_θ is a deep model with parameters $\theta \in \mathbb{R}^N$. To sparsify the deep model, a binary mask $\mathbf{m}$ is often applied through element-wise product $\mathbf{m} \circ \theta$. Such a mask $\mathbf{m}$ is either learned through optimization [205, 204, 206, 184, 185], or obtained based on certain criteria, such as connection sensitivity [200], weight value [201], fisher information [202], or even random initialization [180, 207]. By setting the sparsity ratio $R = 1 - \frac{\|\mathbf{m}\|_0}{|\theta|}$, we can decide how many parameters are rejected from sparse training. Then, we start from a pre-trained model with an initialized mask $\mathbf{m}$.

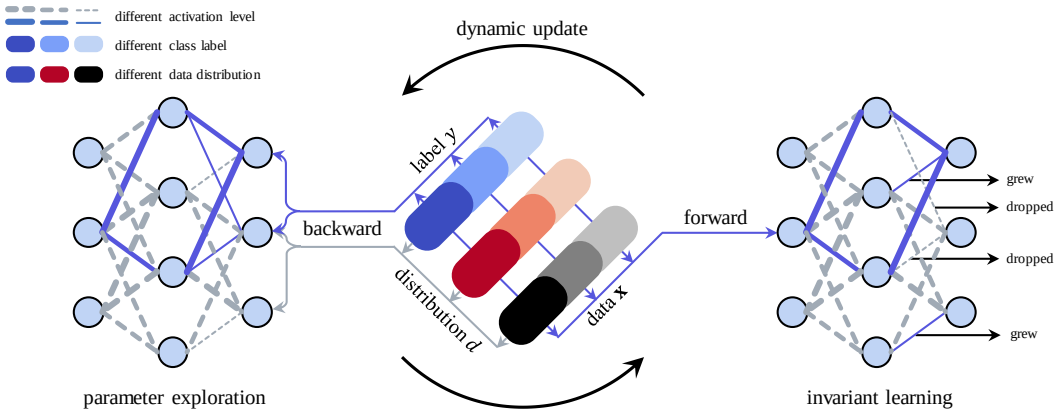


FIGURE 4.2. Learning flow of EVIL: The blocks in the middle are the training dataset where different levels of shades denote different classes, and different types of color indicate different data distributions. The blue arrows (→) and gray arrows (→) stand for the information flow related to label and distribution, respectively. Moreover, the blue solid lines (—) and gray dashed lines (- - -) that connect neurons are the selected invariant parameters and pruned variant ones, respectively.

Parameter Exploration. In this step, we mainly have two optimization targets:

$$\min_{f_{\theta_{inv}} \otimes h} \mathcal{L}_{inv}(h(f_{\theta_{inv}}(\mathbf{x})), y), \tag{4.3}$$

$$\min_{f_{\theta_{var}} \otimes g} \mathcal{L}_{var}(g(f_{\theta_{var}}(\mathbf{x})), d), \tag{4.4}$$

where $\theta_{inv} = \mathbf{m} \circ \theta$ and $\theta_{var} = (1 - \mathbf{m}) \circ \theta$ denote the corresponding invariant parameters and variant ones divided by the mask $\mathbf{m}$, h and g are two fully-connected layers which map the extracted features into class label space $\mathbb{R}^c$ and distribution index space $\mathbb{R}^m$, respectively. Intuitively, the objective $\mathcal{L}_{inv}$ is the classification task, which tries to make predictions based on the label information, and $\mathcal{L}_{var}$ tries to discriminate each distribution based on the distribution information, which is spurious and unwanted.

By minimizing $\mathcal{L}_{inv}$ in Equation (4.3), the gradient magnitude $\nabla_{\theta_{inv}} \mathcal{L}_{inv}$ [200] can be used to find the most relevant parameters to our loss function (Note that other aforementioned sparse training criteria can be used). Similarly, by minimizing $\mathcal{L}_{var}$ in Equation (4.4), those parameters with large $\nabla_{\theta_{var}} \mathcal{L}_{var}$ are sensitive to the spurious information, which cannot help produce invariant features. Thus, we can sort the parameters based on the gradient magnitude to show how much they are activated by their corresponding objective.

Algorithm 3 EVIL

Require: Multiple training sets $\mathcal{E} = \{e_1, \dots e_m\}$; Learning model f_θ; Cosine annealing function $S(t, \alpha, T)$, mask $\mathbf{m}$ initialized based on weight value, iteration number of pre-training T_{pre}.

1: **for** $t = 0, 1, \dots, T-1$ **do**
2: Optimize via Equation (4.3); {*Invariant learning*}
3: **if** $t > T_{\text{pre}}$ and $t\%\Delta T == 0$ **then**
4: Obtain gradients of θ_{inv} and θ_{var} via Equation (4.3) and Equation (4.4), respectively; {*Parameter exploration*}
5: Update the mask $\mathbf{m}$ via Equation (4.5);
6: **end if**
7: **end for**

Further, to dynamically improve our sparsification. We propose to update the mask $\mathbf{m}$ for every ΔT iterations by rejecting the least activated invariant parameters, meanwhile calling back the least activated variant parameters as invariant ones. Specifically, this process is conducted as:

$$\begin{aligned} \mathbf{m}\left[ArgTopK(-|\nabla_{\theta_{inv}}\mathcal{L}_{inv}|, \|\mathbf{m}\|_0 S(t,\alpha,T))\right] &= 0, \\ \mathbf{m}\left[ArgTopK(-|\nabla_{\theta_{var}}\mathcal{L}_{var}|, \|\mathbf{m}\|_0 S(t,\alpha,T))\right] &= 1, \end{aligned} \tag{4.5}$$

where $ArgTopK(v, k)$ returns the indices of top-k elements regarding value v, $\mathbf{m}\left[\cdot\right]$ denotes indexing $\mathbf{m}$. Moreover, to decide how many parameters should be exchanged, we follow Dettmers & Zettlemoyer [204] to use cosine annealing function $S(t, \alpha, T) = \frac{\alpha}{2}(1 + cos(\frac{t\pi}{T}))$, where t and T are the current iteration and total iterations, respectively, and hyper-parameter α decides the largest value. Intuitively, such a cosine annealing function gradually changes from $\alpha < 1$ to 0. Through Equation (4.5), the obtained new mask finds the parameters that are less affected by the distribution information and more related to our learning task than the previous one, further improving the invariant learning performance.

Invariant Learning. After obtaining the updated mask $\mathbf{m}$, we then use the invariant parameters as a subnetwork to conduct invariant learning, which is generally formed as:

$$\mathcal{L}_{inv}(h(f_{\theta_{inv}}(\mathbf{x})), y) = \mathcal{L}_{ce} + \lambda \mathcal{L}_{reg}, \tag{4.6}$$

where the first term is the empirical risk computed through cross-entropy loss, and the second term is the invariant learning regularization with penalty weight λ, which can be realized by

many popular methods. For instance, to use IRM [69], the regularization term is

$$\mathcal{L}_{reg} = \frac{1}{mn} \sum_{\mathbf{x} \in \mathcal{E}} \|\nabla_{h|h=\mathbf{1}} \mathcal{L}_{ce}(h(f_{\theta_{inv}}(\mathbf{x})), y)\|^2. \tag{4.7}$$

To implement REx [144], we penalize the loss variance as

$$\mathcal{L}_{reg} = \frac{1}{mn} \sum_{\mathbf{x} \in \mathcal{E}} Var(\{\mathcal{L}_{ce}(h(f_{\theta_{inv}}(\mathbf{x}|d)), y)\}_{d=1}^{m}). \tag{4.8}$$

Moreover, we can focus on the worst-case distribution to realize DRO [164, 125]:

$$\mathcal{L}_{inv} = \min_{\theta_{inv}} \max_{e \in \mathcal{E}} \frac{1}{n} \sum_{\mathbf{x} \in e} \mathcal{L}_{ce}(h(f_{\theta_{inv}}(\mathbf{x})), y). \tag{4.9}$$

By combining with existing methods, their performance can be largely improved by EVIL, as shown in Section 4.5. The general process of EVIL is summarized in Algorithm 3. Next, we describe one realization of EVIL by adopting the SAM optimizer [137] to further improve the generalization performance.

4.5 Experiment

In this section, we conduct extensive experiments to evaluate the performance of EVIL based on a well-known testbed for OOD generalization: DomainBed [177]. Specifically, we first describe the experimental setup. Then, we improve the performance of well-known invariant learning methods by deploying EVIL, including ERM, IRM [69], REx [144], DRO [164, 125], SAM [137], CORelation ALignment (CORAL) [208], SWAD [139], and MIRO [209]. Further, we compare EVIL and its variant EVIL-SAM with other existing sparse invariant learning methods, including MRM [184], SparseIRM [185], and report the results under different sparsity levels (20%, 40%, 60%, and 80%). Finally, we perform various analytical experiments to validate the effectiveness and efficiency of EVIL.

4.5.1 Practical Implementation

All our experiments are conducted on a single NVIDIA 3090 using PyTorch. To implement our EVIL framework, we first pre-train the models using ERM for 1,000 iterations. Then, a mask $\mathbf{m}$ is initialized based on the weight value. Specifically, by setting a sparsity ratio R, we can select parameters $R|\theta|$-largest weight values by setting their corresponding mask value as 1. During parameter exploration, we first pass the gradient of invariant learning loss $\mathcal{L}_{inv}$, based on which we can sort the invariant parameters with their gradient magnitude from largest to smallest. Then, we reject the $S(t, \alpha, T)$-least invariant parameters by setting their corresponding mask to 0. Similarly, we use the gradient of $\mathcal{L}_{var}$ to sort the variant parameters and recollect top-$S(t, \alpha, T)$ parameters. During invariant learning, we can apply the mask to parameter values as well as their corresponding gradients to conduct sparse training.

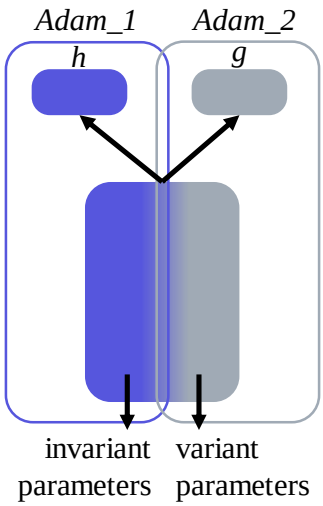


FIGURE 4.3. Illustration of our optimizers.

To optimize our model, we use Adam [210] optimizer with an initial learning rate $1e-3$ without weight decay. Moreover, to avoid the conflict between optimizing invariant parameters and variant parameters, we adopt two Adam optimizers, denoted as $Adam_1$ and $Adam_2$, to correspondingly include the invariant and variant parameters. Moreover, $Adam_1$ would include the class prediction head h and $Adam_2$ would include the distribution prediction head g, as illustrated in Figure 4.3. During the training, $Adam_1$ is mainly used to optimize the invariant parameters, but $Adam_2$ is just employed to optimize the variant ones.

For implementing baseline methods, we mainly follow [177] to set the hyper-parameters. Specifically, for DRO, we set η as $1e-2$ to update the group importance. For IRM, we set $\lambda = 1e2$ to trade off the invariant regularizer. The similar λ for penalization from REx is set to $1e1$. For CORAL and MMD, set the trade-off weight γ as 1. For implementing SagNet, we set the weight for adversarial loss as 0.1. For SAM, we do not use Adaptive SAM [211] and set the perturbation magnitude ρ as 0.05.

TABLE 4.1. Comparison between OOD generalization methods and our EVIL realization on some typical methods. Test accuracies on seven OOD generalization benchmarks from DomainBed. Best results and second best results are highlighted. † denotes results from [139].

Algorithm	**CMNIST**	**RMNIST**	**PACS**	**VLCS**	**OfficeHome**	**TerraInc**	**DomainNet**	**Average**	**FLOPs**
I-Mixup†	31.3	97.8	84.6	77.4	68.1	47.9	39.2	63.7	–
MLDG†	36.9	98.0	84.9	77.2	66.8	47.8	41.2	64.6	–
MMD†	**42.6**	98.1	84.7	77.5	66.4	42.2	23.4	62.1	–
DANN†	29.0	89.1	83.7	78.6	65.9	46.7	38.3	61.6	–
CDANN†	31.1	96.3	82.6	77.5	65.7	45.8	38.3	62.5	–
MTL†	30.4	97.2	84.6	77.2	66.4	45.6	40.6	63.1	–
SagNet†	34.2	96.4	86.3	77.8	68.1	48.6	40.3	64.5	–
ARM†	32.6	98.1	85.1	77.6	64.8	45.5	35.5	62.5	–
RSC†	35.2	96.3	85.2	77.1	65.5	46.6	38.9	63.5	–
Mixstyle†	38.5	97.2	85.2	77.9	60.4	44.0	34.0	62.4	–
ERM†	34.2	98.0	83.3	76.8	67.3	46.2	40.8	63.8	1×
EVIL	39.4 (±1.1)	98.4 (±0.1)	86.0 (±0.1)	78.8 (±0.2)	68.2 (±0.2)	49.1 (±0.2)	43.8 (±0.3)	66.2 (↑ **2.4**)	0.42×
DRO†	32.2	97.9	84.4	76.7	66.0	43.2	33.3	61.9	1×
EVIL-DRO	34.2 (±1.7)	98.2 (±0.1)	85.6 (±0.2)	77.7 (±0.2)	66.4 (±0.1)	49.1 (±0.2)	35.5 (±0.2)	63.8 (↑ **1.9**)	0.42×
IRM†	36.3	97.7	83.5	78.6	64.3	47.6	33.9	63.1	1×
EVIL-IRM	39.1 (±2.2)	98.3 (±0.2)	85.1 (±0.1)	78.8 (±0.1)	66.4 (±0.1)	48.3 (±0.3)	36.0 (±0.3)	64.6 (↑ **1.5**)	0.42×
REx†	39.2	97.3	84.9	78.3	66.4	46.4	33.6	63.7	1×
EVIL-REx	41.2 (±1.3)	98.7 (±0.1)	86.0 (±0.1)	79.1 (±0.2)	68.0 (±0.2)	48.4 (±0.3)	34.5 (±0.1)	65.1 (↑ **1.4**)	0.42×
CORAL†	29.9	98.1	86.2	78.8	68.7	47.7	41.5	64.4	1×
EVIL-CORAL	34.5 (±1.9)	98.6 (±0.1)	86.9 (±0.2)	79.2 (±0.1)	69.0 (±0.1)	49.2 (±0.2)	42.6 (±0.3)	65.7 (↑ **1.3**)	0.43×
SWAD	38.3	98.1	88.1	79.1	70.6	50.0	46.5	67.2	1×
EVIL-SWAD	38.7 (±2.3)	98.3 (±0.3)	**88.3** (±0.1)	79.3 (±0.1)	**71.7** (±0.2)	**51.2** (±0.3)	**46.9** (±0.2)	**67.7** (↑ **0.5**)	0.43×
MIRO	39.4	97.5	85.4	79.0	70.5	50.4	44.3	66.6	1×
EVIL-MIRO	40.2 (±2.3)	98.6 (±0.3)	85.8 (±0.1)	79.4 (±0.1)	71.2 (±0.2)	50.9 (±0.3)	45.0 (±0.2)	67.3 (↑ **0.7**)	0.45×
SAM	38.5	98.1	85.8	79.4	69.6	43.3	44.3	65.6	2×
EVIL-SAM	40.4 (±2.3)	**98.8** (±0.3)	87.8 (±0.1)	**80.1** (±0.1)	70.3 (±0.2)	50.5 (±0.3)	45.0 (±0.2)	67.5 (↑ **1.9**)	0.89×

4.5.2 Experimental Setup

Evaluation Protocol. We follow the experimental setting of DomainBed [177] to evaluation OOD generalization performance. Specifically, DomainBed contains seven benchmark datasets: CMNIST [69] (60,000 images, 10 classes, and 3 domains), RMNIST [212] (60,000 images, 10 classes, and 6 domains), PACS [178] (9,991 images, 7 classes, 4 domains), VLCS [213] (10,729 images, 5 classes, and 4 domains), OfficeHome [214] (15,588 images, 65 classes, and 4 domains), TerraIncognita [215] (24,788 images, 10 classes, and 4 domains), DomainNet [148] (586,575 images, 345 classes, and 6 domains), WILDS [216] (a testbed contains various dataset with significant distribution shfit, here we use two typical datasets: iWildCam and FMoW), ImageNet [104] (contain 1000 classes, here we use ImageNet dataset for fine-tuning, and use many of its variant dataset for OOD evaluation, including: ImageNetV2 [217], ImageNetR [218], ImageNetA [219], ImageNetSketch [220], and ObjectNet [221]). For each benchmark dataset, we leave one domain out of the training dataset and use it as an OOD test dataset. Moreover, we use pre-trained ResNet-50 [3] as our backbone model and train it for 5,000 iterations on all datasets except DomainNet, which requires 15,000 iterations to converge. The test accuracies generated by training models from the last step are provided. To avoid randomness, three independent trials are conducted.

4.5.3 Improving Invariant Learning Using EVIL

In this section, we deploy our EVIL framework to some well-known invariant learning methods and compare them with some other typical baseline methods. To conduct a fair comparison, we only considered end-to-end training on one single model, so some other methods that conduct model ensembling or averaging [139, 222, 223, 224] are not considered. Moreover, we use floating point operations per second (FLOPs) as a criterion to denote the computational efficiency by denoting the FLOPs of ERM as $1\times$ $(7.8e10)$. Practically, we set the sparsity ratio $R = 60\%$, hyper-parameter $\alpha = 0.2$, and $\Delta T = 300$ to implement EVIL. The results are shown in Table 4.1. We can see that our EVIL can effectively improve the performance of all chosen backbone methods. Particularly, on ERM, DRO, and SAM, EVIL can increase their test accuracies for 2.4%, 1.9%, and 1.9%, respectively. Moreover,

TABLE 4.2. Comparison between existing sparse invariant learning methods and EVIL varying sparsity levels. The test accuracies on seven OOD generalization benchmarks from DomainBed are provided. We highlight the **best results** and the second best results.

R	Algorithm	CMNIST	RMNIST	PACS	VLCS	OfficeHome	TerraInc	DomainNet	Average	FLOPs
20%	MRM	31.5	89.3	78.3	70.0	63.6	42.7	37.9	59.0	0.82×
	SparseIRM	31.5	92.2	80.8	71.2	63.7	43.0	39.6	60.3	0.81×
	EVIL	34.1 (±1.7)	95.8 (±0.1)	82.2 (±0.2)	73.7 (±0.2)	66.3 (±0.2)	45.3 (±0.3)	41.9 (±0.2)	62.7	0.82×
	EVIL-SAM	35.2 (±2.3)	96.3 (±0.2)	83.6 (±0.2)	74.0 (±0.2)	65.6 (±0.2)	45.3 (±0.1)	42.6 (±0.1)	63.2	1.62×
40%	MRM	36.2	95.8	81.9	73.5	63.1	45.6	40.4	62.3	0.62×
	SparseIRM	35.7	96.4	82.5	74.2	66.8	47.8	42.6	63.7	0.62×
	EVIL	38.9 (±1.6)	97.3 (±0.2)	84.7 (±0.1)	75.3 (±0.2)	66.4 (±0.1)	47.1 (±0.1)	44.0 (±0.1)	64.8	0.62×
	EVIL-SAM	38.8 (±2.4)	97.9 (±0.3)	84.8 (±0.3)	77.4 (±0.2)	66.9 (±0.1)	48.1 (±0.2)	**45.2** (±0.2)	65.6	1.33×
60%	MRM	38.2	97.6	83.6	76.8	66.5	46.7	40.3	64.2	0.41×
	SparseIRM	37.9	97.9	84.9	77.3	65.1	48.8	42.0	64.8	0.42×
	EVIL	39.4 (±1.4)	98.4 (±0.2)	86.0 (±0.1)	78.8 (±0.2)	68.2 (±0.2)	49.1 (±0.2)	43.8 (±0.3)	66.3	0.42×
	EVIL-SAM	**40.4** (±2.2)	**98.8** (±0.1)	**87.8** (±0.1)	**80.1** (±0.1)	**70.3** (±0.2)	**50.5** (±0.1)	45.0 (±0.2)	**67.6**	0.89×
80%	MRM	37.7	96.3	80.3	72.0	61.2	42.7	35.4	60.8	0.21×
	SparseIRM	37.8	97.2	82.9	71.6	62.4	43.8	36.2	61.7	0.21×
	EVIL	38.5 (±1.3)	98.1 (±0.2)	84.7 (±0.0)	74.1 (±0.2)	64.3 (±0.2)	46.4 (±0.1)	40.1 (±0.0)	63.7	0.21×
	EVIL-SAM	38.9 (±1.2)	98.3 (±0.3)	**87.8** (±0.2)	76.8 (±0.2)	65.7 (±0.2)	47.6 (±0.1)	42.7 (±0.3)	65.4	0.57×

EVIL-SAM achieves the best OOD generalization performance among all compared methods. Especially on TerraIncognita dataset, EVIL-SAM can improve the original performance of SAM for 7.2%, which indicates the effectiveness of EVIL in improving the performance of invariant learning. Moreover, compared to the FLOPs of all baseline methods, our EVIL shows much less computational burden, which manifests the great efficiency of our method.

4.5.4 Comparing EVIL to Sparse Invariant Learning

Furthermore, to show that our method finds a more robust subnetwork for OOD generalization, we compare EVIL with two existing sparse invariant learning methods. Specifically, to validate the effectiveness of our method under different levels of sparsity, we vary to

TABLE 4.3. Results on additional invariant learning methods.

Method	MMD	SagNet	Mixstyle	ARM
w/o EVIL	84.7	86.3	85.2	85.1
with EVIL	**85.3** ($\pm$0.1)	**87.1** ($\pm$0.2)	**86.5** ($\pm$0.2)	**86.6** ($\pm$0.2)

sparsity ratio to 20%, 40%, 60%, and 80%. The experimental results are shown in Table 4.2. Generally, we can see that our EVIL and EVIL-SAM surpass both MRM and SparseIRM in almost all scenarios. Moreover, among all sparsity levels, both EVIL and the other two methods achieve the best results under sparsity 60%. Specifically, EVIL-SAM shows the best performance under sparsity 60% on almost all datasets, and it surpasses the second-best opponent for 2.8% on average accuracy. Besides, EVIL implementation with just ERM can also improve the second-best methods for 1.5% on the averaged results. As for the computational efficiency, our EVIL is comparable to other sparse training methods, except EVIL-SAM, which requires an extra backward pass to compute the parameter perturbation. Therefore, by exploring the variant parameters, EVIL successfully achieves superior OOD generalization performance with comparable efficiency to the sparse invariant learning methods.

4.5.5 Performance on Additional Invariant Learning Methods

We have discussed several invariant learning methods in the main paper, here we conduct extra experiments on PACS dataset using additional invariant learning methods to show how EVIL affects their OOD generalization results. Moreover, we conduct experiments using different network architectures to show the effect of EVIL on various learning models.

Concretely, as we have provided results of IRM, REx, DRO, CORAL, and SAM in the main paper, here we implement EVIL using backbone methods including MMD, SagNet, Mixstyle, and ARM. The results on PACS dataset are shown in Table 4.3. We can see that our method can still improve the OOD generalization performance which is consistent with the observation in the main paper. Therefore, the proposed EVIL framework is generally effective among various invariant learning methods, which shows great deployment practicality.

TABLE 4.4. Results on various model architectures. ResNet50 is pre-trained on ImageNet, and other models are trained from scratch.

Arch.	ResNet50	WRN-20	WRN-32	WRN-44	WRN-56	WRN-110
ERM	84.2	35.6	39.2	41.0	44.6	48.9
EVIL	**86.0** ($\pm$0.1)	**37.3** ($\pm$0.2)	**42.5** ($\pm$0.3)	**43.7** ($\pm$0.3)	**47.2** ($\pm$0.2)	**51.4** ($\pm$0.3)

4.5.6 Performance on Additional ResNet Architectures

Moreover, to evaluate the effectiveness of EVIL on different backbone models, we implement the Wide ResNet (WRN) [101] with varied depths (20, 32, 44, 56, and 110) and train each model from scratch for 500,000 steps to ensure convergence. We also show the result of using ResNet50 pre-trained on ImageNet (Note that due to the pre-training, the performance on ResNet50 would be much better than training from scratch). The comparison between ERM and EVIL is shown in Table 4.4. Again, we can observe the superiority of EVIL over the baseline method ERM on all investigated architectures. Therefore, we can conclude that the performance improvement brought by EVIL is model-agnostic.

4.5.7 Optimizing EVIL Using SAM

In this section, we first briefly describe the realization of EVIL optimized by SAM for OOD generalization (EVIL-SAM). Then, despite of orthogonality of flatness and OOD generalization as found before [139, 222], we discuss some properties of SAM and demonstrate why combining EVIL and SAM can achieve great performance.

Realization of EVIL-SAM. Generally, our EVIL can be optimized using SAM by minimizing the following objectives:

$$\min_{\theta_{inv}} \max_{\|\boldsymbol{\epsilon}\circ\mathbf{m}\|_2\leq\rho} \mathcal{L}(\theta_{inv}+\boldsymbol{\epsilon}\circ\mathbf{m};x,y). \tag{4.10}$$

Specifically, SAM seeks to compute an optimal parameter perturbation $\boldsymbol{\epsilon}^*=\arg\max_{\boldsymbol{\epsilon}}\mathcal{L}(\theta+\boldsymbol{\epsilon};x,y)$ within ρ-radius neighbor that can maximally increase the loss value $\mathcal{L}$. By applying $\boldsymbol{\epsilon}^*$, the loss change $\mathcal{L}(\theta+\boldsymbol{\epsilon}^*;x,y)-\mathcal{L}(\theta;x,y)$ is denoted as *sharpness* which indicates the

TABLE 4.5. Comparison of SAM [137] and ERM under both ID and OOD situations on DomainBed.

		PACS	VLCS	OfficeHome	TerraInc	DomainNet	Average
ID	ERM	96.6	84.7	78.9	91.3	81.4	86.5
	SAM	97.1	86.8	82.0	93.1	85.2	88.8
OOD	ERM	85.5	77.5	66.5	46.1	40.9	63.3
	SAM	85.8	79.4	69.6	43.3	44.3	64.5

flatness of the learned loss function. Intuitively, a flatter loss function often shows better generalization properties, as a slight shift imposed in the input space would not significantly change the loss value. Therefore, SAM has achieved promising in-distribution (ID) generalization performance [225, 159, 211, 160, 170]. To adopt SAM into EVIL, we just need to apply our mask $\mathbf{m}$ to the parameter perturbation $\boldsymbol{\epsilon}$ before computing the optimal $\boldsymbol{\epsilon}^*$. This process not only leaves out spurious information but also reduces the computational burden of SAM. As a result, SAM-EVIL can achieve low sharpness for invariant learning.

Discussion. Although SAM has achieved promising ID results, its OOD performance is quite limited [139, 222] which is still unexplained. As shown in Table 4.5, in the ID scenario, SAM shows great effectiveness compared to ERM, but it merely achieves comparable results to ERM in the OOD setting, even worse in some scenarios. In our perspective, the limitation of SAM is caused by erroneously perturbing the variant parameters which encourages fitting to spurious features. Specifically, in OOD problems, the invariant features and spurious ones would activate θ_{inv} and θ_{var}, respectively. Enforcing robustness (*i.e.*, low sharpness) against perturbation on θ_{inv} can enhance extracting invariant features. However, by perturbing θ_{var}, low sharpness $\mathcal{L}(\theta+\boldsymbol{\epsilon}^*;x,y)-\mathcal{L}(\theta;x,y)$ denotes encouraging the spurious features to bond with the label information. Therefore, SAM cannot extract invariant features as it is sensitive to spurious ones, thus damaging the OOD generalization results. Fortunately, our EVIL can perfectly solve this problem by filtering out the variant parameters which is strongly related to distribution noise. Thus SAM can be further leveraged to enhance the robustness of extracting invariant features. Its effectiveness and efficiency are demonstrated in Section 4.5.

TABLE 4.6. Comparison of EVIL-SAM with other baseline methods on five datasets from DomainNet.

Method	PACS	VLCS	OfficeHome	TerraInc	DomainNet	Avg.
SagNet-SAM	86.4	78.5	69.2	49.3	40.0	64.6
CORAL-SAM	86.6	79.0	69.3	47.9	42.1	65.0
MRM-SAM	83.9	77.1	67.0	47.4	40.6	63.2
SparseIRM-SAM	85.2	77.4	65.6	48.5	43.1	63.9
EVIL-SAM	**87.8**	**80.1**	**70.3**	**50.5**	**45.0**	**66.7**

TABLE 4.7. Performance on ImageNet, iWildCam, and FMoW using CLIP ViT-B/16 as backbone.

Methods	ImageNet		iWildCam		FMoW	
	ID	OOD	ID	OOD	ID	OOD
Zeroshot	68.3 ±0.0	58.7 ±0.0	8.7 ±0.0	11.0 ±0.0	20.4 ±0.0	18.7 ±0.0
Finetuning	82.5 ±0.1	61.3 ±0.1	48.1 ±0.5	35.0 ±0.5	68.5 ±0.1	39.2 ±0.7
EVIL	81.8 ±0.2	**62.5** ±0.6	47.6 ±0.8	**37.4**±1.2	68.2 ±0.6	**41.2** ±1.3

Compare with Other Methods using SAM optimization. To further validate our realization that combining EVIL with SAM indeed shows a positive effect, we compare EVIL-SAM to other algorithms as shown in Table 4.6. We observe that EVIL-SAM achieves the best result among both dense and sparse methods with a significant margin, therefore we can justify our improvement on SAM as more effective than other methods.

4.5.8 Performance on Large-Scale Architecture and Datasets

In this section, we adopt pretrained CLIP ViT-B/16 [226] and conduct finetuning on training datasets from iWildCam, FMoW, and ImageNet, and further test the OOD generalization performance on the split OOD datasets. To extend our sparse training strategy into the CLIP model, we employ a linear layer on top of the ViT backbone and conduct the same pruning strategy by leveraging both class information and domain information. For all datasets, we set the finetuning epoch as 20 and keep the rest of the training parameters the same as described before. The results are shown in Table 4.7, we can see that although EVIL shows a slight performance drop on ID datasets, which is reasonable since we use fewer parameters

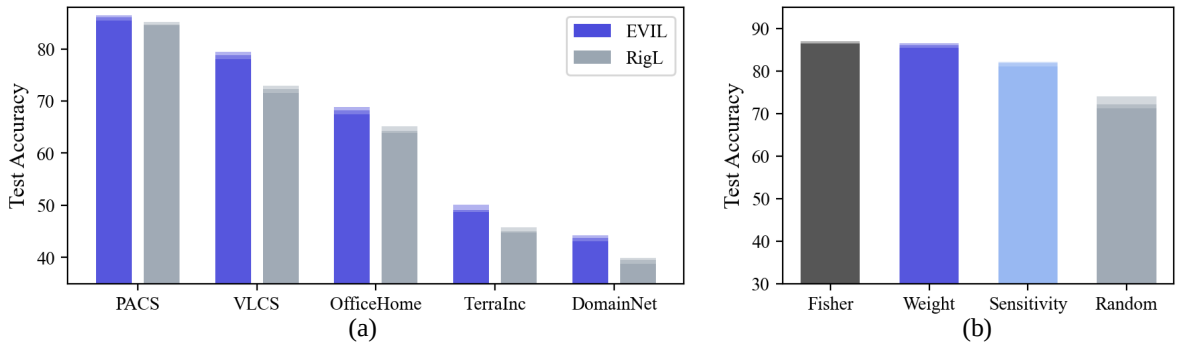


FIGURE 4.4. (a) Comparison of EVIL and RigL which leverages the label information to explore the variant parameters. (b) Comparison of different mask initialization strategies.

than full finetuning, our method achieves the best OOD performance on all three datasets. Specifically, there are 1.2%, 2.4%, and 2.0% performance gains on ImageNet, iWildCam, and FMoW datasets, respectively. Therefore, the effectiveness and superiority of the EVIL can be successfully extended to large-scale architectures and datasets.

4.5.9 Analytical Studies

In this section, we conduct extensive empirical analyses to exploit why EVIL can achieve effective results. First, we conduct ablation studies to show the effect of leveraging distribution knowledge and the influence of choosing different mask initialization strategies. Then, we conduct parameter sensitivity analysis on the hyperparameter α and ΔT. Further, we show a comparison of EVIL-SAM and SAM by visualizing their sharpness during training. Finally, we show the Hessian spectrum to explain why EVIL achieves effective generalization.

Ablation Study. To validate the effectiveness of exploring variant parameters using distribution knowledge, we change the optimization target $\mathcal{L}_{var}(g(f_{\theta_{var}}(\mathbf{x})), d)$ in Equation (4.4) to $\mathcal{L}_{ce}(h(f_{\theta_{var}}(\mathbf{x})), y)$ which leverages the label information instead. As a result, the changed variant is actually Rigging the Lottery (RigL) [201], which is an effective sparse training method. By comparing EVIL and RigL on DomainBed as shown in Figure 4.4 (a), we can see that EVIL surpasses RigL in all scenarios. Therefore, we can conclude that exploring variant parameters by using the distribution information is essential for OOD generalization.

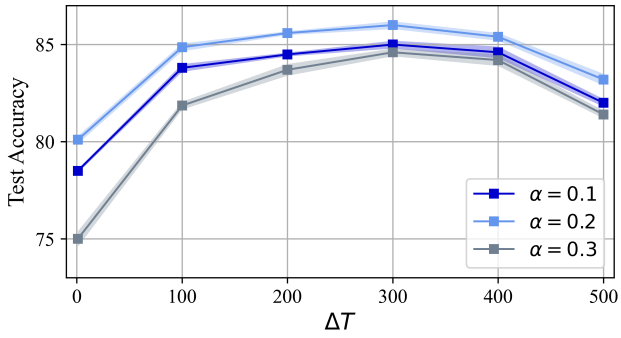


FIGURE 4.5. Parameter sensitivity analysis on α and ΔT.

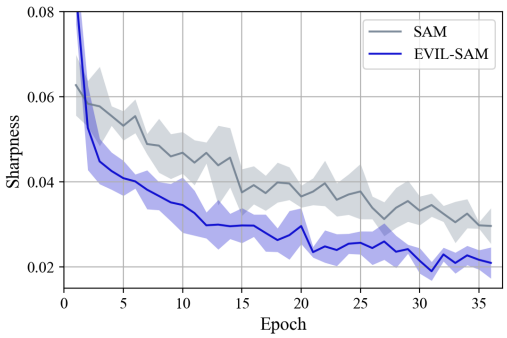


FIGURE 4.6. Sharpness comparison.

Moreover, to show the influence of choosing different mask initialization strategies as mentioned in Section 4.4.1, we compare the weight value (as done in our method) with Fisher information [202], connection sensitivity [200], and random initialization [180] and show the result in Figure 4.4 (b). As we can see, the Fisher information and weight value are two better strategies than connection sensitivity and random initialization, which supports our choice of using the weight value.

Parameter Sensitivity Analysis. To analyze the different choices of α and ΔT, we set α to 0.1, 0.2, and 0.3 and vary ΔT to 1, 100, 200, 300, 400, and 500. As shown in Figure 4.5, choosing α as 0.2, and ΔT as 300 is the best. Moreover, a lower α would hinder the dynamic update of the mask, and a higher α would cause incorporation of noises, thus both choices lead to a performance drop. On the other hand, ΔT controls the updating frequency. Both too small ΔT and too large ΔT would correspondingly cause insufficient update and redundant update, further leading to degradation.

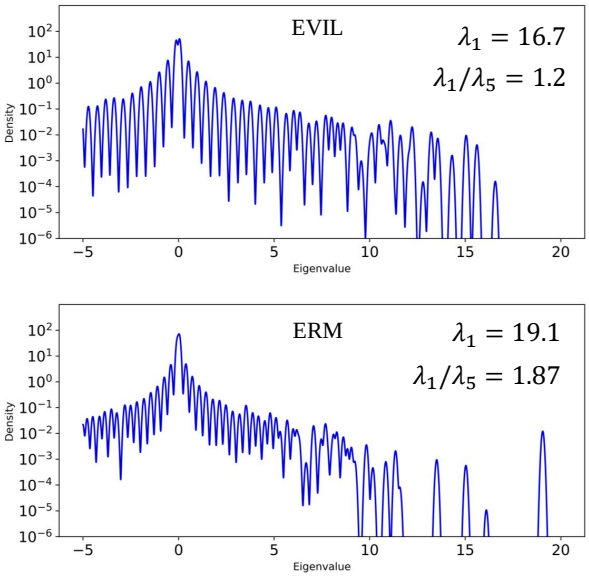
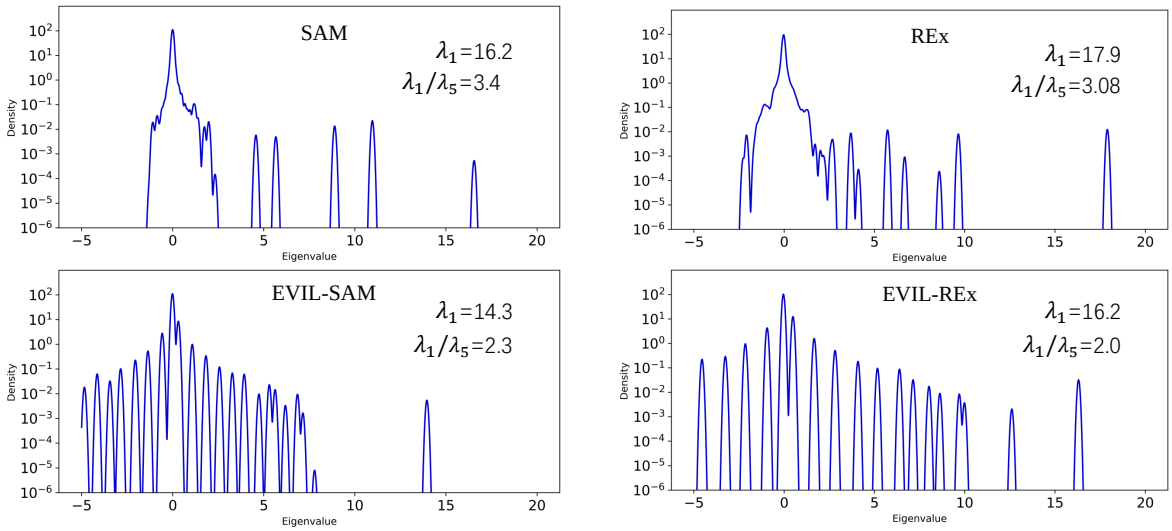


FIGURE 4.7. Hessian Spectrum of EVIL and ERM.



(A) SAM vs. EVIL-SAM (B) REx vs. EVIL-REx

FIGURE 4.8. Hessian spectrum comparison: EVIL realization significantly improves the sharpness across different methods.

Sharpness Analysis. To show how our EVIL affects SAM during OOD generalization, we visualize the sharpness obtained from the training process in Figure 4.6. As a result, our EVIL-SAM can produce smaller sharpness during training than SAM, which indicates that EVIL-SAM is more robust than SAM in OOD generalization problems.

Hessian Spectrum. To analyze whether an algorithm can converge to a flat minima, the Hessian spectrum is commonly used as a criterion [227]. Specifically, we follow Foret et al. [137] to use the ratio of dominant eigenvalue to fifth largest eigenvalue, i.e., λ_1/λ_5 as the criterion for comparing EVIL and ERM. Generally, a smaller λ_1/λ_5 often means a flatter minima is found. Thus, we follow Ghorbani et al. [227] by using the Lanczos algorithm to approximate the Hessian spectrum of ERM and EVIL in Figure 4.7. As we can see, the λ_1/λ_5 of EVIL is much smaller than that of ERM, which confirms that our method can converge to a flatter minima than ERM. Moreover, as the dominant eigenvalue λ_1 is also an important measurement, we can see that EVIL produces a smaller λ_1 than ERM as well, which again supports the effectiveness of EVIL. Therefore, it is reasonable that EVIL can achieve great generalization results.

Additional Hessian Spectrum on SAM and REx. Since the proposed method shows effective generalization performance, as we have demonstrated in Section 4.5.9, we further validate that the proposed EVIL framework can still help produce improved Hessian spectrum when compared to other methods such as SAM and REx. As shown in Figs. 4.8a and 4.8b, We observe the same phenomenon as in the main paper: when combined with EVIL, the largest eigenvalue of both SAM and REx is smaller than its original ones, and the Hessian spectrum are more compact when using our EVIL framework. Therefore, we can again conclude that EVIL indeed helps produce flat minima.

4.6 Conclusion

In this Chapter, we aim to address the problem that existing sparse invariant learning methods fail to fully capture invariant information in OOD generalization problems, owing to the misleading influence of distribution shifts. Therefore, we propose EVIL by leveraging the distribution knowledge to explore the variant parameters. By finding the variant parameters that are highly sensitive to distribution shift, we can identify a robust subnetwork that effectively extracts invariant features. Moreover, we propose to improve our identification dynamically during network training. As a result, our EVIL framework can effectively and efficiently improve the OOD generalization performance of many invariant learning methods, meanwhile surpassing all compared sparse invariant learning methods. Exhaustive analyses are conducted to comprehensively validate the performance of EVIL.

CHAPTER 5

Machine Vision Therapy

Although pre-trained models such as Contrastive Language-Image Pre-Training (CLIP) show impressive generalization results, their robustness is still limited under Out-of-Distribution (OOD) scenarios. Instead of undesirably leveraging human annotation as commonly done, it is possible to leverage the visual understanding power of Multi-modal Large Language Models (MLLMs). However, MLLMs struggle with vision problems due to task incompatibility, thus hindering their effectiveness. In this Chapter, we propose to effectively leverage MLLMs via Machine Vision Therapy, which aims to rectify erroneous predictions of specific vision models. By supervising vision models using MLLM predictions, visual robustness can be boosted in a nearly unsupervised manner. Moreover, we propose a Denoising In-Context Learning (DICL) strategy to solve the incompatibility issue. Concretely, by examining the noise probability of each example through a transition matrix, we construct an instruction containing a correct exemplar and a probable erroneous one, which enables MLLMs to detect and rectify the incorrect predictions of vision models. Under mild assumptions, we theoretically show that our DICL method is guaranteed to find the ground truth. Through extensive experiments on various OOD datasets, our method demonstrates powerful capabilities for enhancing visual robustness under many OOD scenarios.

5.1 Introduction

Pre-trained vision models such as Vision Transformers (ViT) [228, 229, 230] with Contrastive Language-Image Pretraining (CLIP) [231, 226, 232, 233, 234, 235] have been widely

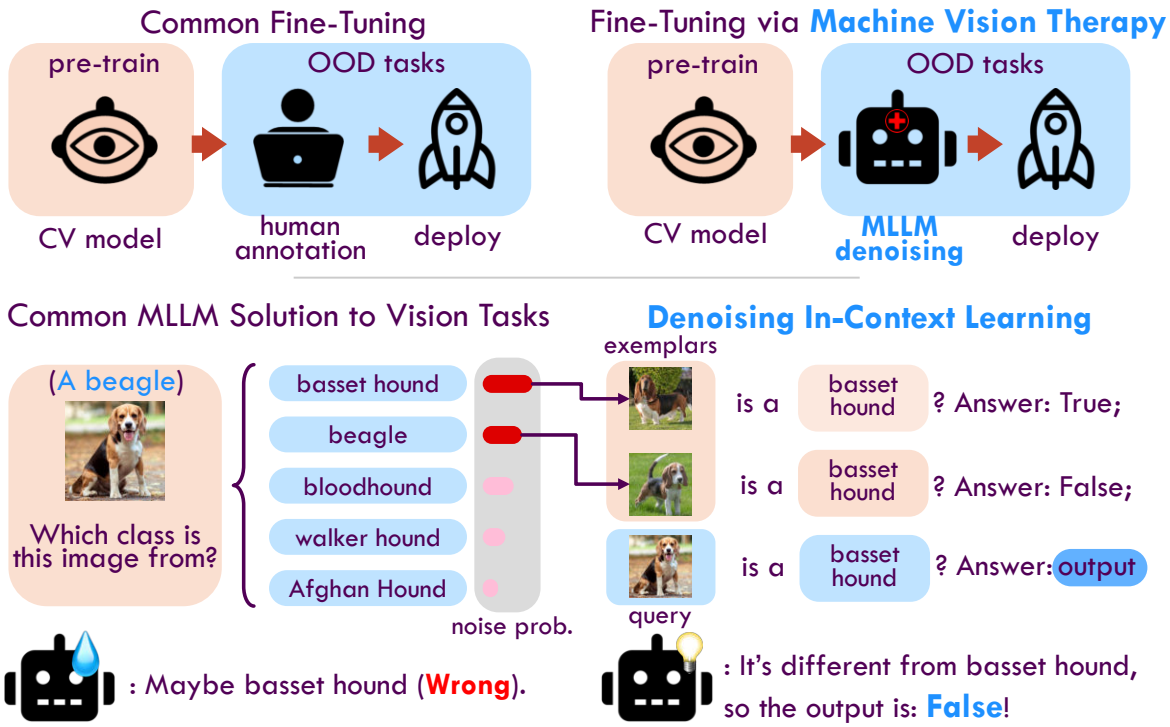


FIGURE 5.1. Illustration of our methodology: Upper row: Comparison between common fine-tuning process and fine-tuning via Machine Vision Therapy. Our method potentially eliminates the necessity for human-annotation by leveraging the knowledge from MLLMs. Lower row: Comparison between previous MLLM solution to vision tasks and Denoising In-Context Learning strategy. Instead of considering all classes, our method make predictions by presenting a pair of positive and negative exemplars.

used thanks to their strong generalization performance meanwhile effectively avoiding training vision models from scratch. But when deployed to Out-of-Distribution (OOD) scenarios [236, 237, 28, 238, 239, 46, 240, 220, 241], their recognition performance could be seriously degraded [242]. Downstream fine-tuning has been a common practice to regain the generalizability [243, 244, 245], but it requires additional label acquisition through human labor, which is undesirable for large-scale applications.

Fortunately, the thriving Multi-modal Large Language Models (MLLMs) [246, 247, 248, 249, 250, 251, 252, 253, 254], which take advantage of the few-shot learning ability of Large Language Models (LLM) [255, 256, 257, 258, 259, 260, 261, 262, 263], have manifested powerful capabilities on understanding visual information with language interpretations, and excelled at recognizing novel objects in multimodal tasks such as image captioning, visual

question answering, visual reasoning, etc. Considering the vulnerability of vision models under OOD situations, here we hope to refine vision models by leveraging the knowledge of MLLMs, as shown in the upper row of Figure 5.1. However, due to the difficulty of aligning the text generation process with visual recognition tasks[1] [246, 264], MLLMs struggle with generating correct answers that match the ground-truth class names, thus underperforming the current dominant contrastive paradigms, even when employing them as own vision encoders [246, 247, 265, 264, 266].

Focusing on enhancing the robustness of vision models, in this Chapter, we propose to effectively leverage MLLMs to conduct **Machine Vision Therapy (MVT)** which aims to diagnose and rectify the error predictions through a novel Denoising In-Context Learning (DICL) strategy. Then, we utilize the rectified supervision to guide the fine-tuning process in downstream OOD problems. Specifically, rather than giving a set of options to ask MLLMs for the exact answer [246, 265, 266], we show that it is sufficient to query for the ground truth by using only two exemplars, i.e., (1) a correct one that demonstrates the exact match between a query class name with its image example and (2) an erroneous one that combines the same query class with an image from the most confusing category for the vision model. Since the erroneous predictions are essentially label noise, hence we draw inspiration from learning with noisy labels [267, 268, 269, 270, 271, 272, 273, 123, 274, 275, 64, 276, 277]. Particularly, we can find the erroneous categories by estimating a transition matrix that captures the probability of one class being mistaken as another. By feeding the two exemplars, MLLMs can be instructed to leverage their few-shot learning power to distinguish the semantically similar images that are easily misclassified by vision models, as shown in the lower row of Figure 5.1. To process such instructions, we leverage the multi-modal in-context learning ability of several existing MLLMs [278, 279, 280, 281] to realize our methodology. After the error predictions are diagnosed and rectified, vision models can be further fine-tuned to enhance their OOD robustness on downstream data distribution. Through a comprehensive empirical study on many challenging datasets and their OOD variants, such as ImageNet [282], WILDS [216], and DomainBed [177], we carefully validate

[1]Here, we mainly focus on classification task.

the effectiveness of MVT and demonstrate its superiority under various OOD scenarios on many well-known vision models.

To sum up, our contributions are threefold:

- We design a novel Machine Vision Therapy paradigm to enhance computer vision models by effectively leveraging the knowledge of MLLMs without needing additional label information.
- We propose a Denoising In-Context Learning strategy to successfully align MLLMs with vision tasks.
- Through comprehensive quantitative and qualitative studies on many well-known datasets, we demonstrate that the proposed method can enhance: (1) generalization on both ID and OOD data, (2) robustness against domain shift, (3) robustness against common corruptions, (4) performance on recognizing fine-Grained attributes, (5) robustness against spurious correlations, (6) detection on prediction errors and OOD data.

5.2 Related Work

In this section, we provide a brief discussion of the OOD generalization problem and multimodal large-language models.

5.2.1 OOD Generalization

OOD data refers to those with different distributions from training data. OOD generalization aims at improving the performance of deep models to unseen test environments. Researchers attempted to tackle the problem from different perspectives, such as data augmentation, OOD detection, invariant causal mechanisms [153, 283, 284, 285], and so on. Data augmentation is effective in improving model generalization. Typical methods involve Cutout [286], which randomly occludes parts of an input image; CutMix [287], which replaces a part of the target image with a different image; Mixup [288], which produces a convex combination

of two images; DeepAugment [218], which passes a clean image through an image-to-image network and introduces several perturbations during the forward pass. Some methods conduct OOD detection to separate OOD data. Typical methods include softmax confidence score [28, 189], which is a baseline for OOD detection; Outlier Exposure (OE) [289], which uses unlabeled data as auxiliary OOD training data. Energy scores are shown to be better for distinguishing OOD samples from IID ones [36]. Some work resort to causality to study the OOD generalization problem. Typical methods include MatchDG [151], which proposes matching-based algorithms when base objects are observed and approximate the objective when objects are not observed; INVRAT [142], which leveraged some conditional independence relationships induced by the common causal mechanism assumption.

5.2.2 Multimodal Large Language Models

The field of vision-language models has witnessed significant advancements in recent years, driven by the growing synergy between computer vision and natural language processing. Notably, this synergy has led to the exceptional zero-shot performance [290] of CLIP [226], a model that employs a two-tower contrastive pretraining approach to align image and text information. In the rapidly evolving landscape of LLMs, such as GPTs [255], LLaMA [260], and Vicuna [291], it has become evident that LLMs possess the capacity to process information from diverse domains. BLIP-2 [250], for instance, serves as a foundational model, aligning visual features and text features using a Querying Transformer (Q-former) and utilizing OPT [292] and FLAN [256] as language models. Building upon BLIP-2, Instruct-BLIP [293] has enhanced instruction-following capabilities. To further bolster the instruction-following proficiency of multi-modal models, LLaVA [252] and Mini-GPT4 [254] have introduced meticulously constructed instruction sets, which have found widespread application in various multi-modal models. mPLUG-Owl [253] introduces a two-stage learning paradigm, first fine-tuning the visual encoder and then refining the language model with LoRA [294]. This approach effectively fuses image and text features. Some models consider additional modalities, such as ImageBind [295], which simultaneously incorporates data from six modalities without the need for explicit supervision, and PandaGPT [296] which enhances its

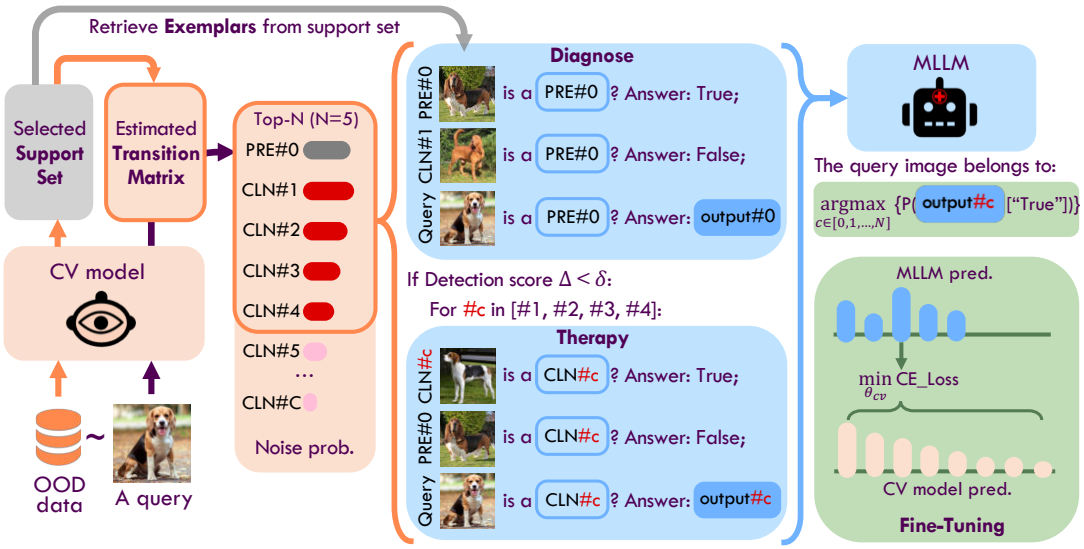


FIGURE 5.2. Workflow of our Machine Vision Therapy: The orange part demonstrates the Transition Matrix Estimation, the blue part indicates the Denoising In-Context Learning process, and the green part illustrates the Fine-Tuning of vision models.

instruction-following capabilities. Several multi-modal models prioritize the in-context learning abilities of LLMs. Flamingo [246], in one of the pioneering efforts, integrates a gated cross-attention module to align with the spaces of images and text. Otter [251] refines OpenFlamingo [247], an open-source version of Flamingo, improves instruction-following abilities. Multi-Modal In-Context Learning (MMICL) [281] is a comprehensive vision-language model that incorporates Instruct-BLIP, enabling the analysis and comprehension of multiple images, as well as the execution of instructions. MLLMs possess the remarkable capacity to capture intricate details and engage in reasoning when presented with an image. Nevertheless, it remains uncertain about how to enhance visual perception by harnessing the knowledge embedded within LLMs.

5.3 Methodology

In this section, we carefully demonstrate the Machine Vision Therapy process which mainly contains three components, namely Transition Matrix Estimation, Denoising In-Context

Learning, and Fine-Tuning of vision models. Next, we demonstrate problem setting and framework overview.

5.3.1 Problem Formulation and Overview

Generalizing to Out-of-Distribution tasks has been a challenging topic in computer vision problems, where we normally have a vision model parameterized by $\theta_{cv} \in \Theta_{cv}$ pre-trained on massive labeled in-distribution (ID) data $\mathcal{D}^{id} = \{x_i^{id}, y_i^{id}\}_{i=0}^{m} \in \mathcal{X} \times \mathcal{Y}$, where $\mathcal{Y} = \mathbb{R}^C$. Here each ID example is sampled from a joint distribution, i.e., $(X^{id}, Y^{id}) \sim p^{id}$, where X^{id} and Y^{id} stand for variables. After pretraining, we can assume the conditional distribution $P(Y^{id}|X^{id})$ can be perfectly captured by the inference function $\tilde{y}^{id} = f_{\theta_{cv}}(x^{id})$, where $\tilde{y}^{id}$ is the prediction. In OOD tasks, we are given a set of unlabeled examples $\mathcal{D}^{ood} = \{x_i^{ood}\}_{i=0}^{n}$ whose element $x^{ood} \in \mathcal{X}$ is drawn from an unknown data distribution p^{ood}. Due to the change of downstream task, some factors that affect the data generating process are shifted, causing a difference between p^{ood} and p^{id}, further hindering the label prediction, i.e., $\tilde{y}^{ood} = f_{\theta_{cv}}(x^{ood}) \not\sim P(Y^{ood}|X^{ood})$, where Y^{ood} is the unknown ground truth. Fortunately, having been observed with extraordinary low-shot generalization capability, we leverage MLLM with parameters $\theta_{mllm} \in \Theta_{mllm}$ to enhance the OOD robustness of vision models.

Our framework is illustrated in Figure 5.2 and our problem can be formulated as follows:

$$\begin{aligned} \min_{\theta_{cv}} \mathcal{L}(f_{\theta_{cv}}, z);\ z = \big[\theta_{mllm}((X_c^+, Y_c^+); (X_c^-, Y_c^+); X_i)\big]_c^N; \\ Y_c = T[c; \operatorname{argmax}[f_{\theta_{cv}}(X_i)]], \end{aligned} \tag{5.1}$$

where X_i^+ and X_i^- denotes the positive and negative exemplars, respectively, X_i is the query image, and T is the transition matrix. Intuitively, when a distribution shift occurs, the emerging prediction errors are essentially label noises that can be captured by estimating a transition matrix. Hence, by focusing on calibrating the examples with high noise probabilities, the visual robustness of downstream tasks can be improved effectively. In particular, we feed all OOD data into the vision model to obtain the noisy prediction distribution $P(\tilde{Y}^{ood}|X^{ood})$,

based on which we can effectively estimate T and provide exemplars to instruct MLLM[2]. Further, we conduct machine vision therapy to find the possible ground truth for X_i based on the MLLM output z. Finally, z is leveraged to minimize $\mathcal{L}$ to optimize θ_{cv}. Next, we explain the details of each process.

5.3.2 Transition Matrix Estimation

The distribution shift from OOD data x^{ood} leads to unreliable label prediction $\tilde{y}^{ood}$, which is highly unreliable due to instance-dependent feature noises [297, 123] as shown in Section 5.4.8. Hence, in order to capture the relationship between $\tilde{Y}^{ood}$ and Y^{ood}, we leverage a transition matrix $T \in [0,1]^{C\times C}$ [268, 271, 122] which satisfies $P(Y^{ood}|X^{ood}) = T^{\top}P(\tilde{Y}^{ood}|X^{ood})$. However, estimating such a transition matrix is difficult without access to any noisy label supervision or strong assumption [268, 122]. Therefore, we propose a simple yet effective sample selection approach to construct a support set with clean labels. Specifically, we rank all OOD data within each class based on their prediction confidence, i.e., $\max_c \left[f_{\theta_{cv}}(x^{ood})\right]_c$, where $[\cdot]_c$ denotes the value of the c-th entry. From the sorted dataset $\{x_1^{ood,c}, x_2^{ood,c}, \cdots, x_{\frac{n}{C}}^{ood,c}\}_{c=1}^{C}$, we uniformly sample ρ examples per class, where ρ is the labeling budget, i.e., $\mathcal{D}^{supp} = \{\{x_{j\times\frac{n}{\rho C}}^{ood,c}\}_{j=1}^{\rho}\}_{c=1}^{C}$. In this way, we can effectively model the noisy posterior $P(\tilde{Y}^{ood}|X^{ood})$. Then, through an acceptable labeling process[3], we can obtain the clean label posterior $P(Y^{ood}|X^{ood})$, thus effectively estimating the transition matrix T. Finally, the noise transition probability $T[:, \arg\max[f_{\theta_{cv}}]]$ of a query image can be obtained by indexing T through its current prediction.

[2]Although some manual annotation is required, we show in later experiments that our strategy has an acceptable labeling workload and demonstrates superior performance to vanilla fine-tuning on the support set. Furthermore, the support set is **not used for parameter tuning** in our method, so our fine-tuning does not actually use any human annotation for training.

[3]We experimentally show that when there is a distribution shift between $\mathcal{D}^{supp}$ and $\mathcal{D}^{ood}$, the proposed method can still perform effectively. As a result, it is unnecessary to conduct the labeling process on each practical task. Instead, we can just use the existing support set to instruct most of OOD tasks.

5.3.3 Denoising In-Context Learning

Thanks to the previously obtained noise probability list $T\,[:, \arg\max[f_{\theta_{cv}}]]$, we can further decide which one is the possible ground truth through DICL. In particular, we only consider the classes of the top-N noise probability as potential candidates. If the label prediction denoted by "PRE#0" is not in the candidates, we would fix it in the first place. Further, we conduct *Diagnosing* which decides the fidelity of the current prediction, and *Therapy* which finds the possible ground truth.

Diagnosing. Since the inference time of MLLMs is non-trivial, it is necessary to avoid redundant analysis on confident examples. Hence, to examine the fidelity of vision model predictions, our Diagnosing focuses on answering whether a query image belonging to class "PRE#0" is "True". Specifically, we retrieve from $\mathcal{D}^{supp}$ to obtain one exemplar image belonging to "PRE#0", and another exemplar image belonging to the class with the largest noise transition probability "CLN#1"[4]. Then, combined with the query image X_q, an in-context instruction is constructed:

> Question: This image <IMG_PRE#0> shows a photo of <PRE#0>, True or False?
> Answer: True;
> Question: This image <IMG_CLN#1> shows a photo of <PRE#0>, True or False?
> Answer: False;
> Question: This image <IMG_Query> shows a photo of <PRE#0>, True or False?
> Answer:

The symbols <IMG_PRE#0>, <IMG_CLN#1>, and <IMG_Query> are replace tokens for the image features of exemplars from "PRE#0" and "CLN#1", and X_q, respectively. The first exemplar acts as the positive one to show MLLMs the true image from class "PRE#0", and the second exemplar shows the negative one to show the highly probable false image from "CLN#1". Then, based on the X_q and "PRE#0", MLLMs can effectively judge the

[4]The performance of retrieve strategy is carefully studied in Section 5.4.9.

correctness by outputting z_0:

$$z_0 = \theta_{mllm}((X_{PRE\#0}, Y_{PRE\#0}); (X_{CLN\#1}, Y_{PRE\#0}); X_q). \tag{5.2}$$

To enable further quantitative analysis, we obtain the logits of "True" and "False" tokens from the MLLM output z_0 followed by a softmax function:

$$z_0 := \text{softmax}([z_0[\text{True}], z_0[\text{False}]]). \tag{5.3}$$

Finally, we combine $z_0[\text{True}]$ and the prediction confidence of the vision model to obtain a detection score Δ:

$$\Delta = \frac{1}{2}(z_0[\text{True}] + \max_c [f_{\theta_{cv}}]_c (x^{ood})). \tag{5.4}$$

If Δ is larger than a threshold δ, we assume the current prediction "PRE#0" is correct[5], otherwise, we conduct the next Therapy process.

Therapy. During therapy, we continue to use the instruction template above and traverse across the rest clean class candidates. Particularly, for each iteration c in $N-1$ trials, we choose "CLN#c" as the positive class and "PRE#0" as the negative class, whose exemplars are correspondingly retrieved from $\mathcal{D}^{supp}$ to construct the prompt. Then, it is fed into MLLM to output whether the query image belongs to the class "CLN#c", i.e., $z_c = \theta_{mllm}((X_{CLN\#c}, Y_{CLN\#c}); (X_{PRE\#0}, Y_{CLN\#c}); X_q)$,let $z_c := \text{softmax}([z_c[\text{True}], z_c[\text{False}]])$. As a result, we can decide the final prediction through:

$$y_{mllm} = \arg\max [z_c[\text{True}]]_{c=0}^{N}. \tag{5.5}$$

As shown in Section 5.4, the performance of MLLM prediction shows strong performance in many OOD scenarios. However, we still cannot directly employ MLLMs for inference, due to three main reasons: (1) Non-negligible inference time: Since current MLLMs cannot handle large-batch data, it would be unimaginably slower (*e.g.*, 1000×) when using MLLMs rather than vision models; (2) High requirements for computation: Inference through MLLM takes up huge memory of GPU. For MLLMs using large LLMs such as LLaMA-13B, it requires distributed inference on less advanced devices; (3) Model privacy issue: Many

[5]Detailed analysis is shown in Section 5.4.9.

Algorithm 4 Machine Vision Therapy.

Require: Pre-trained vision model θ_{cv}, MLLM θ_{mllm}, OOD dataset $\mathcal{D}^{ood}$.

1: Uniformly sample ρC examples from confidence-sorted $\mathcal{D}^{ood}$ to construct support set $\mathcal{D}^{supp}$;
2: Estimate transition matrix T; {*Section 5.3.2*}
3: **for** $i = 0, 1, \cdots, n$ **do**
4: Based on label prediction $\tilde{y}_i^{ood}$, obtain noisy transition probability $T\,[:, \arg\max[f_{\theta_{cv}}]]$;
5: Conduct Diagnosing through Equation (5.2) and compute detection score Δ through Equation (5.4);
6: **if** $\Delta > \delta$ **then**
7: Accept current prediction;
8: **else**
9: Conduct Therapy and obtain MLLM prediction through Equation (5.5); {*Section 5.3.3*}
10: Based on the MLLM prediction, conduct fine-tuning through Equation (5.6); {*Section 5.3.4*}
11: **end if**
12: **end for**

MLLMs are highly sensitive with limited accessibility, therefore. Hence, we propose to fine-tune vision models based on the prediction of MLLMs.

5.3.4 Fine-Tuning of Vision Models

After obtaining the MLLM prediction y_{mllm}, we propose to optimize vision models through the following objective:

$$min_{\theta_{cv}} \mathcal{L}_{ce}(f_{\theta_{cv}}, y_{mllm}), \tag{5.6}$$

where $\mathcal{L}_{ce}(\cdot)$ denotes the cross-entropy loss. Here we summarize our methodology in Algorithm 4. Further, we can directly deploy the fine-tuned vision models to OOD tasks whose effectiveness is demonstrated in Section 5.4.

5.3.5 Theoretical Analysis

We denote the MLLM is pretrained over a distribution p defined by a latent concept $\phi \in \Phi$. During DICL, there are n examples to form a prompt S_n which are sampled from a prompt distribution p_{prompt} defined by concept $\phi^* \in \Phi$. To justify the proposed DICL strategy,

based on the theoretical framework proposed by Xie et al. [298], we show that when MLLM achieving the most probable z based on the given prompt S_n and query image-text pair x_q-y under a concept ϕ^*, the corresponding y is the same as the one found from p_{prompt}, which is y_q that matches with x_q.

ASSUMPTION 3 (Distribution consistency). $\forall(x_q, y_q) \sim p_{prompt}, p(x_q, y_q) = p_{prompt}(x_q, y_q)$.

Moreover, the assumptions from Xie et al. [298] also hold, then we have:

THEOREM 4. *Assume that the above assumptions hold, if for all $\phi \in \Phi$, $\phi \neq \phi^*$, the concept ϕ^* satisfies the distinguishability condition: $\sum_{j=1}^{k} KL_j(\phi^* \| \phi) > \epsilon_{start}^{\phi} + \epsilon_{delim}^{\phi}$, then as $n \to \infty$, the prediction according to the pretraining distribution is*

$$\arg\max_y p(y|S_n, x_q, \phi^*) \to \arg\max_y p_{prompt}(y|x_q). \tag{5.7}$$

Thus, the in-context predictor f_n achieves the optimal $0-1$ risk: $\lim_{n\to\infty} \mathcal{L}_{0-1}(f_n) = \inf_f \mathcal{L}_{0-1}(f)$.

LEMMA 1. *Under the same condiction of Theorem 4, the prediction z according to the pretraining distribution is*

$$\operatorname{argmax}_z p(z|S_n, x_q, y_q, \phi^*) \to \operatorname{argmax}_z p_{prompt}(z|x_q, y_q). \tag{5.8}$$

THEOREM 5. *Assume that the above assumptions hold, as $n \to \infty$, when achieving the largest prediction probability of z given prompt under concept ϕ^*, the corresponding class description y follows the same y obtained from the prompt distribution:*

$$\operatorname{argmax}_y p(z|S_n, x_q, y, \phi^*) \to \operatorname{argmax}_y p_{prompt}(z|x_q, y). \tag{5.9}$$

We can see that if n is large enough, the MLLM prediction z achieves the largest value when y_q is the exact match to x_q. As a result, we can justify that only when we feed the positive

TABLE 5.1. Classification accuracy (%) of baseline CLIP models and our method on 5 ID datasets and 5 OOD datasets. The baseline methods includes ViT-L from CLIP [226] and ViT-g from EVA [299], VQA, and Vanilla FT.

Arch	Method	ID					OOD				
		IN-Val	IN-V2	CIFAR10	CIFAR100	MNIST	IN-A	IN-R	IN-SK	IN-V	iWildCam
RN50	CLIP	59.7	52.6	71.5	41.9	58.5	23.9	60.7	35.4	31.1	8.2
RN101	CLIP	61.7	56.2	80.8	48.8	51.6	30.2	66.7	40.9	35.4	12.3
ViT-B	CLIP	62.9	56.1	89.9	65.0	47.9	32.2	67.9	41.9	30.5	10.9
ViT-L	CLIP	75.8	70.2	95.6	78.2	76.4	69.3	86.6	59.4	51.8	13.4
	VQA	64.9	59.9	97.6	**83.2**	56.7	66.0	87.3	56.9	56.2	13.3
	Vanilla FT	76.1	**70.8**	96.1	80.3	77.5	70.8	87.5	60.0	53.6	15.2
	MVT	75.2	**70.8**	**97.9**	78.9	53.0	71.2	88.1	59.0	62.1	**25.0**
	+FT	**76.9**	70.5	96.7	82.0	**79.2**	**75.1**	**89.5**	**61.4**	**68.8**	-
ViT-g	EVA	78.8	71.2	98.3	88.8	62.2	71.9	91.4	67.7	64.9	21.9
	VQA	64.3	59.6	97.9	84.5	55.7	64.6	87.4	58.2	59.2	19.7
	Vanilla FT	78.9	71.8	98.7	89.1	62.9	72.7	91.6	68.1	65.6	22.4
	MVT	**79.1**	71.6	98.1	89.0	63.2	73.2	91.4	67.9	66.3	**25.1**
	+FT	79.0	**72.2**	**98.9**	**91.2**	**65.7**	**75.5**	**92.8**	**68.6**	**70.6**	-

image-text pair to the MLLM, the prediction z is the largest among all other combinations between x_q and $y \in \mathcal{Y}, y \neq y_q$.

5.4 Experiment

In this section, we first provide our experimental details. Then we conduct quantitative comparisons with the state-of-the-art vision models. Finally, we conduct ablation studies and analyses to qualitatively validate our method.

5.4.1 Experimental Setup

Datasets. In our experiments, we use well-known ID datasets including ImageNet-1K [282] validation dataset, ImageNet-V2 [217], CIFAR10 [107], CIFAR100 [107] and MNIST [300]. We also evaluate OOD generalization on datasets that are commonly considered OOD ones, ImageNet-A [219], ImageNet-R [218], mageNet-Sketch [220], ImageNet-V [301], iWildCam [302], and DomainBed [303].

TABLE 5.2. Classification accuracy (%) of baseline CLIP models and our method on 4 subsets of DomainBed datasets. The baseline methods include ViT-L from CLIP, ViT-g from EVA, VQA, and Vanilla FT.

	Datasets	VLCS				PACS			
	Method	0	1	2	3	0	1	2	3
ViT-L	CLIP	74.9	83.5	80.3	74.5	97.8	97.4	97.5	99.4
	Vanilla FT	78.8	85.2	83.4	77.0	**98.0**	97.6	97.7	99.6
	MVT	83.8	89.0	87.2	80.3	97.6	97.5	**98.0**	99.4
	+FT	**84.2**	**89.8**	**87.9**	**82.5**	**84.2**	**98.2**	**98.0**	**99.8**
ViT-g	EVA	72.5	80.0	79.8	72.8	99.0	98.8	98.9	99.8
	Vanilla FT	75.5	82.3	82.1	75.6	98.9	98.7	98.9	99.8
	MVT	81.2	86.6	86.1	79.5	98.2	98.0	98.0	99.4
	+FT	**83.7**	**89.5**	**86.9**	**82.0**	**99.1**	**98.9**	**99.0**	**100.0**

	Datasets	OfficeHome				DomainNet					Avg
	Method	0	1	2	3	0	1	2	3	4	
ViT-L	CLIP	87.7	92.7	85.7	85.6	61.1	62.1	60.2	78.4	51.1	80.6
	Vanilla FT	87.9	93.1	87.1	86.9	62.0	62.5	60.5	78.5	51.9	81.6
	MVT	87.7	93.4	89.0	88.5	61.3	62.1	60.4	78.7	53.4	82.8
	+FT	**90.9**	**95.0**	**90.9**	**90.8**	**62.5**	**63.8**	**62.4**	**80.1**	**54.0**	**84.0**
ViT-g	EVA	90.5	94.2	88.6	88.7	61.4	64.7	61.2	81.6	54.9	81.6
	Vanilla FT	90.6	94.5	89.2	89.0	61.5	64.9	61.3	81.8	54.8	82.3
	MVT	89.7	93.8	89.7	89.1	**62.2**	**65.0**	61.6	**82.3**	56.1	83.3
	+FT	**91.6**	**95.1**	**90.7**	**90.6**	61.9	64.8	**63.2**	81.9	**56.6**	**84.4**

Models and baselines. For vision backbone, we employ CLIP models [226] and utilize ViT-L/14 and ViT-g [304] from EVA [299] as the vision model to be enhanced. For the MLLM backbone, we consider two existing works MMICL [281] and Otter [251] that possess multimodal ICL ability. Additionally, we conduct Visual Question Answering (VQA) to directly ask MLLMs the class of query images. Moreover, we conduct vanilla fine-tuning (Vanilla FT) using only D^{supp} as a baseline. The performance of using MLLM prediction is denoted as MVT, and our fine-tuning result is denoted as FT.

Settings. For model evaluation, we randomly select 5000 images independently from the ImageNet validation set (IN-Val), ImageNet-V2 (IN-V2), ImageNet-A (IN-A), ImageNet-R (IN-R), ImageNet-Sketch (IN-SK), ImageNet-V (IN-V) and 10000 images independently from CIFAR10, CIFAR100, MNIST to constitute the test samples. Additionally, we select 3

TABLE 5.3. Classification accuracy (%) of baseline CLIP models and our method with MMICL [281] and Otter [251] as the VLMs on 5 ID datasets and 5 OOD datasets. We compare the performance of our method, and the fine-tuned models supervised by our method with the baseline models, i.e., ViT-L from CLIP [226]. Fine-tuning with both MMICL and Otter improves the classification accuracy.

MLLM	Method	ID				
		IN-Val	IN-V2	CIFAR10	CIFAR100	MNIST
None	CLIP	75.8	70.2	95.6	78.2	76.4
MMICL	MVT	75.2	**70.8**	**97.9**	78.9	53.0
	+FT	**76.9**	70.5	96.7	**82.0**	79.2
Otter	MVT	74.2	67.4	94.7	70.1	52.0
	+FT	76.3	70.1	96.6	81.8	**81.3**

MLLM	Method	OOD					Avg (ID+OOD)
		IN-A	IN-R	IN-SK	IN-V	iWildCam	
None	CLIP	69.3	86.6	59.4	51.8	13.4	61.7
MMICL	MVT	71.2	88.1	59.0	62.1	**25.0**	64.3
	+FT	**75.1**	**89.5**	**61.4**	**68.8**	-	**75.6**
Otter	MVT	64.1	85.2	59.5	51.9	16.2	60.3
	+FT	73.5	88.7	60.0	55.7	-	73.0

images per category to construct a support set to provide in-context exemplars. We evaluate iWildCam from WILDS and VLCS, PACS, OfficeHome, and DomainNet from DomainBed. For the details of implementation, we choose the top-6 noisy classes to conduct MVT. Concretely, we set the threshold $\delta = 0.6$ to diagnose incorrect predictions, then we retrieve exemplars from the support set based on the most similar logit prediction to query images. For each round of DICL, we repeat the process for 3 times and average the model predictions. During fine-tuning, we optimize the vision models for 3 epochs using Adam and SGD optimizers for ViT-L and ViT-g, respectively.

5.4.2 Quantitative Comparison

First, we compare our MVT method with well-known vision models under both ID and OOD scenarios. As shown in Table 5.1, we can see that our method with fine-tuning denoted as

TABLE 5.4. Classification accuracy (%) of baseline CLIP models and our method with MMICL [281] and Otter [251] as the VLMs on 4 subsets of DomainBed datasets, including VLCS, PACS, OfficeHome, and DomainNet. We compare the performance of our method and the fine-tuned models supervised by our method with the baseline models, i.e., ViT-L from CLIP [226]. Fine-tuning with both MMICL and Otter improves the classification accuracy.

MLLM	Datasets	VLCS				PACS			
	Method	0	1	2	3	0	1	2	3
None	CLIP	74.9	83.5	80.3	74.5	97.8	97.4	97.5	99.4
MMICL	MVT	83.8	89.0	87.2	80.3	97.6	97.5	**98.0**	99.4
	+FT	**84.2**	**89.8**	**87.9**	**82.5**	**98.0**	**98.2**	**98.0**	**99.8**
Otter	MVT	67.5	77.4	73.7	66.6	97.0	96.3	96.5	99.0
	+FT	76.8	87.7	82.3	77.4	**98.0**	97.7	**98.0**	**99.8**

MLLM	Datasets	OfficeHome				DomainNet					Avg
	Method	0	1	2	3	0	1	2	3	4	
None	CLIP	87.7	92.7	85.7	85.6	61.1	62.1	60.2	78.4	51.1	80.6
MMICL	MVT	87.7	93.4	89.0	88.5	61.3	62.1	60.4	78.7	53.4	82.8
	+FT	**90.9**	**95.0**	**90.9**	**90.8**	**62.5**	**63.8**	**62.4**	**80.1**	**54.0**	**84.0**
Otter	MVT	85.6	89.9	83.6	83.3	56.5	58.6	56.3	74.1	46.5	77.0
	+FT	88.7	93.4	87.7	87.1	62.0	63.0	61.3	79.7	53.4	82.0

"+FT" achieves better performance in most settings. Specifically, on "IN-V", our method with fine-tuning can significantly surpass both CLIP and EVA for 17% and 6%, respectively. Moreover, on "IN-A", our method achieves 4.3% and 2.8% performance improvement over the second-best method on both ViT-L and ViT-g backbone, respectively. We can also observe that even without fine-tuning, the prediction accuracy of MLLM denoted by "MVT" can still surpass all baselines on most scenarios, which denotes the strong performance enhancement of our MVT fine-tuning on vision models. Note that we did not provide fine-tuning on iWildCam because most of the predictions are incorrect. Though MVT can still achieve the best result, the vision encoders could be misled by erroneous decisions during the fine-tuning process.

Furthermore, we consider domain shift by leveraging DomainBed datasets. Specifically, for each dataset, we leave one domain out as a test dataset and fine-tune on rest domains. By

TABLE 5.5. Classification accuracy (%) of baseline CLIP models and our method on 5 ID datasets and 5 OOD datasets. We compare the performance of our method, and the fine-tuned models supervised by our method with the baseline models, including ResNet-50 and ViT-B/32. The supervisor MLLM is MMICL [281].

Arch	Method	ID				
		IN-Val	IN-V2	CIFAR10	CIFAR100	MNIST
RN50	CLIP	59.7	52.6	71.5	41.9	**58.5**
	MVT	**76.2**	**70.8**	**80.2**	**49.7**	50.8
	+FT	66.3	65.7	75.1	46.9	47.3
ViT-B	CLIP	62.9	56.1	89.9	**65.0**	47.9
	MVT	**77.5**	**71.0**	**92.5**	60.4	**51.5**
	+FT	66.3	66.0	90.1	59.5	46.6

Arch	Method	OOD				
		IN-A	IN-R	IN-SK	IN-V	iWildCam
RN50	CLIP	23.9	60.7	35.4	31.1	8.2
	MVT	**47.5**	**72.9**	**41.6**	**54.1**	**14.5**
	+FT	32.1	64.4	36.5	38.2	-
ViT-B	CLIP	32.2	67.9	41.9	30.5	10.9
	MVT	**60.6**	**83.0**	**47.8**	**53.1**	**19.3**
	+FT	38.8	68.7	43.1	37.6	-

comparing two state-of-the-art vision backbones ViT-L and ViT-g, we show the performance comparison in Table 5.2. As we can see, both MVT and MVT with fine-tuning can significantly surpass the baseline methods. For some scenarios such as the PACS dataset, our method can achieve nearly 100% performance. Moreover, in several scenarios in the VLCS dataset, both our MVT and fine-tuning can achieve almost 10% improvements. Additionally, we find that our method with fine-tuning largely surpasses vanilla fine-tuning baseline on both Tables 5.1 and 5.2. Hence, we can conclude that our learning strategy can indeed provide effective supervisions which enhances vision robustness under distribution shift.

5.4.3 Quantitative Comparison using Otter

Similarly, here we conduct additional experiments on various ImageNet-based datasets and DomainBed datasets using ViT-L but a different MLLM backbone: Otter [251]. The results

TABLE 5.6. Classification accuracy (%) of baseline CLIP models and our method on 4 subsets of DomainBed datasets, including VLCS, PACS, OfficeHome, and DomainNet. We compare the performance of our method and the fine-tuned models supervised by our method with the baseline models, including ResNet-50 and ViT-B/32. The supervisor MLLM is MMICL [281].

Arch	Method	VLCS				PACS			
		0	1	2	3	0	1	2	3
RN50	CLIP	75.0	82.3	81.3	75.0	91.3	90.3	90.0	96.2
	MVT	**84.3**	**88.0**	**88.7**	**81.8**	**96.0**	**96.1**	**95.4**	**98.8**
	+FT	83.7	87.3	88.1	81.3	95.6	95.7	95.1	98.6
ViT-B	CLIP	74.0	82.0	79.6	74.4	93.6	92.8	93.0	98.2
	MVT	**84.2**	**87.3**	**88.4**	**82.8**	**96.7**	**96.4**	**96.6**	**98.8**
	+FT	76.0	84.8	81.3	81.6	92.9	88.8	89.4	93.3

Arch	Method	OfficeHome				DomainNet					Avg
		0	1	2	3	0	1	2	3	4	
RN50	CLIP	71.7	80.9	69.4	67.8	47.2	46.8	44.9	64.0	32.9	71.0
	MVT	**77.2**	**85.3**	**77.5**	**75.4**	**46.1**	**46.3**	**43.4**	**61.7**	**33.2**	**75.0**
	+FT	75.9	85.0	75.8	74.7	45.3	45.6	43.0	60.4	32.6	74.3
ViT-B	CLIP	79.2	86.4	77.4	76.3	49.7	54.3	51.0	68.7	40.7	74.8
	MVT	**84.0**	**89.3**	**82.9**	**81.5**	**49.5**	**53.1**	**51.5**	**69.9**	**41.7**	**78.5**
	+FT	81.1	88.3	80.8	77.7	47.2	52.5	48.7	66.9	40.5	74.8

are shown in Tables 5.3 and 5.4. We find that the performance of MVT is dependent on the MLLM backbone: when using Otter as the backbone model for MVT, the OOD performance would slightly degrade from the performance of MMICL, which could be due to the capability of MLLM to conduct ICL. However, the rectified predictions can still contain useful information to boost the performance of vision models. In several cases in ImageNet-Val, MNIST, and ImageNet-R, Otter with fine-tuning can still improve the visual robustness to the best or second-best results.

5.4.4 MVT on Additional Vision Models

Then, we conduct MVT using MMICL but using different vision backbone models such as ViT-B and ResNet-50 on ImageNet and DomainBed datasets. The results are shown in Tables 5.5 and 5.6. We can see that the performance of MVT is quite strong compared to

TABLE 5.7. Classification accuracy (%) of baseline CLIP models and our method with MMICL [281] as the VLM on 15 corruptions and 5 severities of ImageNet-C datasets. We compare the performance of our method and the fine-tuned models supervised by our method with the baseline models, i.e., ViT-L from CLIP [226]. The fine-tuned models with our MVT method have the best performance.

	Gaussian Noise						Shot Noise						Impulse Noise					
	1	2	3	4	5	avg	1	2	3	4	5	avg	1	2	3	4	5	avg
CLIP	69.8	66.7	59.7	46.9	30.6	54.7	70.5	64.9	57.7	43.6	32.1	53.8	65.7	60.2	55.9	45.0	32.7	51.9
MVT	70.1	67.5	61.2	**49.8**	**33.6**	56.4	70.8	66.8	59.2	**46.1**	**35.6**	55.7	66.3	61.9	58.4	47.5	35.6	53.9
+FT	**71.0**	**67.9**	**61.3**	48.7	33.5	**56.5**	**72.0**	**67.1**	**60.1**	**46.1**	35.2	**56.1**	**68.5**	**64.2**	**59.8**	**48.9**	**35.8**	**55.4**
	Defocus Blur						Glass Blur						Motion Blur					
	1	2	3	4	5	avg	1	2	3	4	5	avg	1	2	3	4	5	avg
CLIP	66.1	62.4	53.0	43.4	35.0	52.0	65.5	59.3	40.5	33.8	25.4	44.9	70.9	66.8	59.9	49.5	41.8	57.8
MVT	67.1	63.3	55.8	**47.6**	**38.8**	54.5	67.1	61.3	42.8	36.0	29.4	47.3	71.9	67.7	60.9	51.5	43.2	59.0
+FT	**68.8**	**64.1**	**56.3**	47.4	38.4	**55.0**	**68.9**	**64.8**	**45.2**	**37.6**	**30.2**	**49.3**	**72.8**	**69.1**	**62.1**	**52.7**	**45.3**	**60.4**
	Zoom Blur						Snow						Frost					
	1	2	3	4	5	avg	1	2	3	4	5	avg	1	2	3	4	5	avg
CLIP	62.2	55.9	49.8	43.9	37.3	49.8	68.3	61.2	61.9	56.1	52.6	60.0	68.5	61.2	53.8	51.1	46.6	56.2
MVT	64.1	57.3	52.0	45.7	38.7	51.6	69.2	61.5	62.9	57.1	54.0	60.9	69.5	61.5	54.2	52.7	47.4	57.1
+FT	**65.2**	**59.2**	**54.2**	**48.8**	**41.4**	**53.8**	**70.6**	**63.9**	**64.6**	**59.2**	**55.6**	**62.8**	**71.9**	**65.2**	**57.9**	**56.4**	**51.5**	**60.6**
	Fog						Brightness						Contrast					
	1	2	3	4	5	avg	1	2	3	4	5	avg	1	2	3	4	5	avg
CLIP	69.8	67.9	65.0	61.3	52.0	63.2	74.3	74.0	72.8	70.6	68.1	72.0	70.6	69.3	64.8	52.4	35.1	58.4
MVT	70.7	69.2	66.5	62.6	53.8	64.6	74.7	74.1	72.6	71.1	68.8	72.3	70.9	69.9	65.1	52.9	36.9	59.1
+FT	**72.5**	**71.3**	**69.5**	**67.1**	**60.3**	**68.1**	**76.0**	**75.1**	**74.3**	**73.1**	**71.1**	**73.9**	**73.5**	**73.5**	**70.2**	**59.2**	**42.7**	**63.8**
	Elastic						Pixelate						JPEG					
	1	2	3	4	5	avg	1	2	3	4	5	avg	1	2	3	4	5	avg
CLIP	69.2	50.6	64.1	53.1	30.4	53.5	71.0	70.4	66.2	60.1	54.6	64.5	70.8	67.7	65.1	58.0	45.3	61.4
MVT	70.0	51.1	65.8	55.2	**32.7**	55.0	71.7	70.7	66.5	61.9	57.3	65.6	**72.5**	**69.6**	**67.5**	**60.5**	47.7	**63.6**
+FT	**70.7**	**53.6**	**67.7**	**58.5**	32.2	**56.5**	**72.8**	**71.8**	**69.1**	**62.9**	**57.7**	**66.9**	71.2	68.9	65.8	60.0	**48.8**	62.9

other vision models which shows over 10% and 4% improvements in ImageNet datasets and DomainBed datasets, respectively. Especially on ImageNet-V2, ImageNet-A, ImageNet-R, and ImageNet-V, the performance improvement of MVT are encouragingly over 15%, 24%, 12%, and 23%, respectively. After fine-tuning, the performance can be improved in most cases, such as ResNet-50 is further improved by 13.1% and 3.3% correspondingly on ImageNet-V2 and DomainBed thanks to MMICL.

(a) One-To-One (O2O) (b) Many-To-Many (M2M)

FIGURE 5.3. Figures are from Lynch et al. [305], the letters on each images denote a certain background. There are two spurious correlation types in the Spawrious dataset, namely O2O and M2M. In the O2O setting, each dog class is correlated to one certain background type and different distributions have different correlation probabilities as shown by the bar below the O2O figure. As for the M2M setting, multiple classes and backgrounds are correlated together and the correlation changes to different groups of classes and backgrounds during testing.

5.4.5 Robustness against Visual Corruptions

Further, we consider the visual robustness against corruptions by evaluating EVIL on a robustness benchmark: ImageNet-C [114]. Specifically, there are 15 different types of corruption with different corruption severities varied from 1 to 5. Here we cover all scenarios to evaluate our method using MMICL as a backbone model and a baseline method CLIP ViT-L. The results are shown in Table 5.7. We can see that our method shows very strong performance in all scenarios. Compared to CLIP, using MVT can improve the performance by over 2%, and through fine-tuning, the performance is further boosted by over 4%. The encouraging results again demonstrate the effectiveness of our method.

5.4.6 Robustness against Spurious Correlation

Moreover, we consider a common distribution shift scenario where the training dataset and test dataset have different foreground and background correlation, *i.e.*, spurious correlation. Specifically, as shown in Figure 5.3, standing for the Spawrious dataset that we use, there are

TABLE 5.8. Performance comparison between MVT and CLIP on robustness against spurious correlation using Spawrious dataset.

Type	O2O_easy	O2O_medium	O2O_hard	M2M_easy	M2M_medium	M2M_hard	Avg.
ViT-L	94.1	95.4	93.3	96.7	95.0	92.5	94.5
MVT	**95.8**	**96.3**	**93.6**	**96.8**	**95.8**	**92.9**	**95.2**
ViT-g	94.6	97.0	92.6	96.7	95.6	94.8	95.2
MVT	**95.3**	**97.4**	**92.8**	**96.8**	**96.6**	**95.4**	**95.7**

two different settings: One-To-One (O2O) correlation, where each class is correlated to one background type with a certain probability. The foreground objects in the training dataset and test dataset have different probabilities of being combined with a certain background. For the Many-To-Many (M2M) setting, the foregrounds and backgrounds are split into subgroups that contain multiple classes and background types. When different subgroups are correlated together between training and test datasets, the M2M spurious correlation is formed and brings more complexity. In the Spawrious dataset, there are three levels of hardness based on the correlation probability difference between the training and test datasets, namely easy, medium, and hard. Here, we consider all scenarios and show the results in Table 5.8. We can see that the MVT method can outperform the ViT-L and ViT-g baseline methods in all scenarios, which leads to the conclusion that our method is robust to spurious correlations and can identify the class of interests despite the changing backgrounds.

TABLE 5.9. Class names for 12 chosen attributes.

Attribute	-1	+1
Male	a woman	a man
Wear_Hat	not wearing a hat	wearing a hat
Smiling	not smiling	smiling
Eyeglasses	not wearing eye glasses	wearing eye glasses
Blond_Hair	not having blond hair	having blind hair
Mustache	not having mustache	having mustache
Attractive	not attractive	attractive
Wearing_Lipstick	not wearing lipstick	wearing lipstick
Wearing_Necklace	not wearing necklace	wearing necklace
Wearing_Necktie	not wearing necktie	wearing necktie
Young	not young	young
Bald	not bald	bald

FIGURE 5.4. Examples of celebA photos with different attributes.

5.4.7 Performance on Recognizing Fine-grained Attributes

Additionally, here we further explore the capability of recognizing subtle attributes based on the CelebA dataset [306]. Particularly, we consider 12 face attributes, as shown in Figure 5.4. For each attribute, we testify whether a learning model could correctly identify the attribute in a given image. Here we compare our MVT method with CLIP ViT-L and ViT-g, and the performance of MVT produced by conducting therapy on ViT-L and ViT-g models.

Particularly, since CelebA is a binary classification task, here we design different prompts for vision models and our MLLM. For CLIP models, we use `The person in this image is <#classname>` as text input, where `<#classname>` of each attribute is shown in Table 5.9. For our method, we still designed one positive prompt and one negative prompt for each ICL round. Specifically, for "Male" attribute, our instruction is as follows:

TABLE 5.10. Performance comparison between MVT and CLIP on recognizing fine-grained attributes using CelebA dataset.

Attr.	Male	Hat	Smiling	Glasses	Blond	Mustache	Attract	Lipstick	Necklace	Necktie	Young	Bald	Avg.
ViT-L	63.0	60.8	64.5	75.8	36.2	29.0	42.0	30.8	38.0	37.5	66.6	86.3	52.5
MVT	**74.0**	**67.0**	**65.4**	**76.1**	**53.0**	**55.8**	**42.4**	**39.4**	**38.6**	**53.9**	**73.5**	**88.1**	**60.6**
ViT-g	98.5	75.5	70.4	83.8	46.0	66.6	58.2	72.5	43.5	28.4	54.1	91.3	65.7
MVT	**98.9**	**77.2**	**71.0**	**84.1**	**58.3**	**74.9**	**59.0**	**73.2**	**43.6**	**41.2**	**56.1**	**91.8**	**69.1**

Question: Is the person in this image {replace_roken} a male? Answer: True;

Question: Is the person in this image {replace_roken} a female? Answer: False;

Question: Is the person in this image {replace_roken} a male? Answer:

in which is first exemplar demonstrates an image of a male positively described as male, the second exemplar shows an image of a male negatively described as female, and finally, we ask whether the input image is a male and use the output of MLLM as the prediction.

The results on CelebA are shown in Table 5.10, we observe that our method is quite effective in recognizing fine-grained attributes and its performance significantly surpasses ViT-L and ViT-g with a large margin. Especially in attributes such as "Blond_Hair", "Mustache", and "Wearing_Necktie", the performance improvements are even over 10% on both two CLIP models, and the final averaged results on all 12 attributes, the total improvements are 8.1% and 3.4% for ViT-L and ViT-g, respectively. Therefore, it is reasonable to conclude that our method can be effectively conducted on fine-grained attribute recognition and significantly outperforms several powerful vision models.

5.4.8 Ablation Study

In this part, we conduct ablation studies to analyze each module of MVT by using ViT-L backbone vision model.

Ablation Study on Transition Matrix Estimation. To validate the performance of transition matrix estimation, we compare our confidence-based uniform sampling strategy to a random sampling baseline. The result on the ImageNet-V dataset is shown in Figure 5.5. To

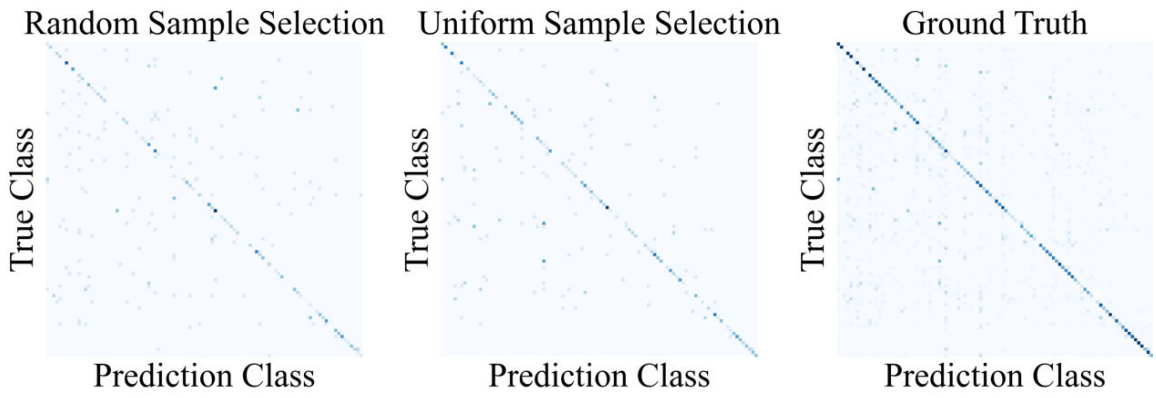


FIGURE 5.5. Ablation study on transition matrix estimation by comparing our method with random sampling and ground truth.

TABLE 5.11. Performance comparison between choosing noisy classes via transition matrix (MVT) and using Top-N predictions.

	IN-A	IN-SK	IN-Val	IN-R	IN-V2	IN-V
Top-N Pred.	60.3	58.4	74.2	85.3	67.7	58.3
MVT	65.5	59.0	75.1	86.0	70.7	61.6

quantitatively show the superiority of our method, we compute the ℓ_2 norm of the difference between one estimation and ground truth which indicates the fidelity of the estimation. As a result, our estimation is much more accurate by achieving 3.83 norm, compared to 4.46 of random sampling.

Ablation Study on Choosing Noisy Classes. Further, we justify the choice of using a transition matrix to obtain the noisy classes. As a comparison, we use the top-6 predictions as the therapy candidates and show the results in Table 5.11. We can see on all datasets, our method can outperform the opponent with non-trivial improvements. Therefore, leveraging the transition matrix to find the potential noisy classes is more effective than using prediction.

Ablation Study on Detection Score. To analyze the proposed detection score on conducting diagnosing, we show the distribution of prediction confidence provided by the vision model, MLLM, and our detection score Δ in Figure 5.6. Based on the results, we can justify our design of Δ: In the left column, we can see the confidence of correctly classified examples is very high, but the wrong ones show uniform distribution. Conversely, in the middle column,

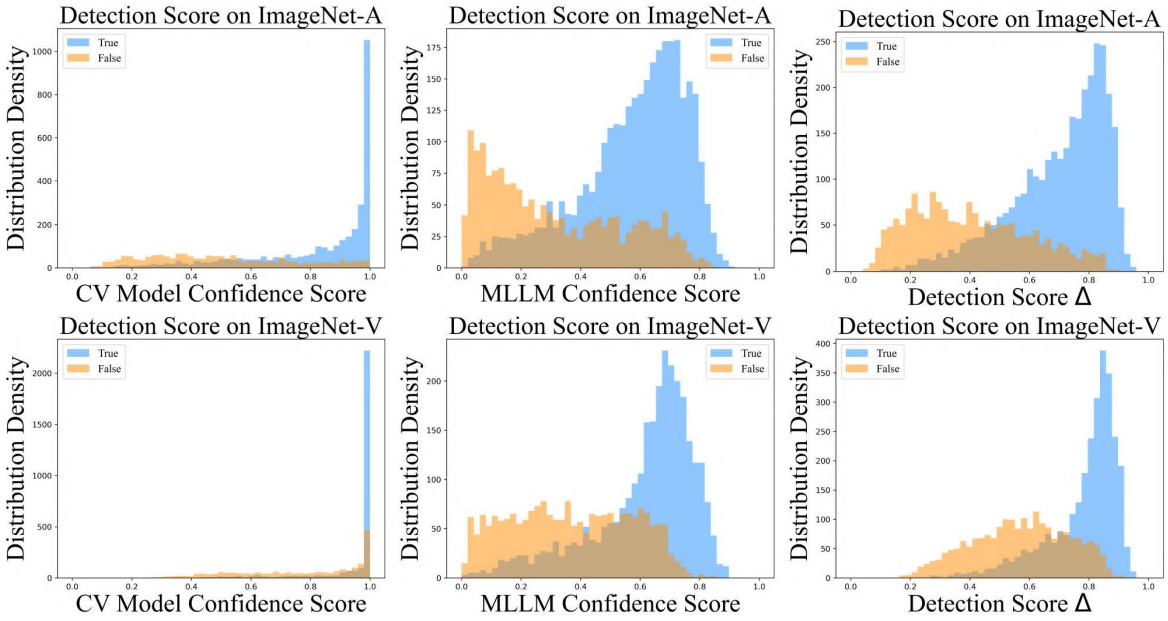


FIGURE 5.6. Ablation study on detection score distribution.

TABLE 5.12. Comparison of classification accuracy (%) on 5 OOD datasets with Otter [251] and MMICL [281]. We compare the performance on CLIP ViT-L [226] backbone.

MLLM	Method	IN-A	IN-R	IN-SK	IN-V	iWildCam
None	CLIP	69.3	86.6	59.4	51.8	13.4
Otter	MVT	64.1	85.2	59.5	51.9	**16.2**
	+FT	**73.5**	**88.7**	**60.0**	**55.7**	-
MMICL	MVT	71.2	88.1	59.0	62.1	**25.0**
	+FT	**75.1**	**89.5**	**61.4**	**68.8**	-

although MLLM poses slightly lower scores on correct ones, it significantly suppresses the confidence of wrong ones. As a result, we combine two scores to obtain Δ, which can produce clearly separable distributions to benefit the diagnosing process. Unless specified, we set the threshold $\delta = 0.6$ which works effectively in most scenarios.

Ablation Study on MLLM Backbone. To testify the effectiveness of MVT on different MLLM backbones, here we instantiate our method using Otter [279] and compare it to the previous realization on MIMIC [281]. The result is shown in Table 5.12. We can see that both the implementation on MMICL and Otter show superior performance to the employed vision

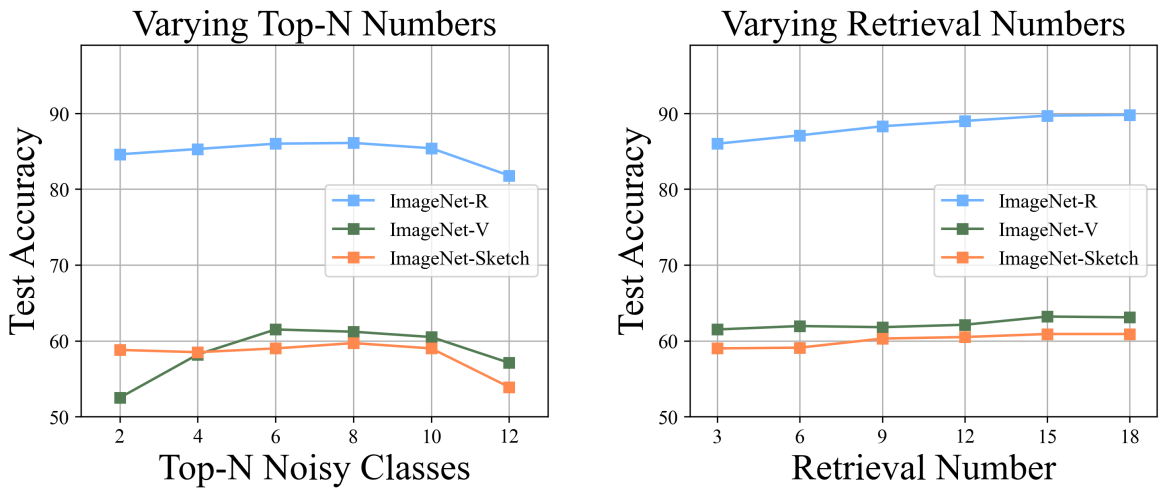


FIGURE 5.7. Performance analysis: (left) varying the number of top-N chosen noisy classes; (right) varying the number of retrieved exemplars.

encoder backbone. Although the performance slightly differs between Otter and MMICL, which could be due to the model capacity and their training strategy, we can generally conclude that our MVT method is applicable to different MLLMs backbones with ICL and could further benefit from more sophisticated MLLMs in the future.

5.4.9 Performance Analysis

Further, we conduct qualitative analysis to thoroughly validate the effectiveness of our MVT.

Choice of Top-N Noisy Classes. To study how a varied number of chosen noisy classes could affect the performance of our method, we change the top-N number from 2 to 12, and show the result on ImageNet-R, ImageNet-V, and ImageNet-Sketch datasets in Figure 5.7 left. We find a common phenomenon that either too small or too large a number of N could hurt the performance. This could be because that small N would ignore too many potential ground-truth classes. In contrast, large N includes too many choices that could interfere with the final prediction. Setting N to 6 could be an ideal choice for ImageNet-based datasets.

Effect of Retrieval Numbers. In our experiments, we retrieve exemplars for 3 times and average the predictions. To further investigate the effect of varied retrieval numbers, we change the number of retrievals from 3 to 18 and conduct experiments on the same OOD datasets

as above. Specifically, we consider one positive and negative pair for a single DICL round as one retrieval. We repeat this process for R times and ensemble the MLLM predictions through $\frac{1}{R}\sum_r^R [z_c[\text{True}]^r, z_c[\text{False}]^r]$. In this way, it is possible that MLLM predictions would be more accurate. The result is shown in Figure 5.7 right. We observe that the performance steadily improves as the retrieval number increases, however, the performance gains vanish when the retrieval number becomes too large. Moreover, large retrieval numbers would multiply the computation cost. Therefore, it is suggested to set the number reasonably small.

Performance of Different Retrieval Strategy. As shown by Alayrac et al. [246], Retrieval-based In-Context Example Selection (RICES) can significantly affect the ICL performance. Therefore, here we investigate its influence. Specifically, we propose two retrieval strategies, namely feature-based retrieval and logit-based retrieval. The former one is based on feature similarity and the latter one is based on the prediction logit. For each strategy, we conduct experiments on selecting the most similar examples and the least similar examples, which are denoted as "most" and "least", respectively. The results are shown in Figure 5.8. Apart from the intuitive finding that least-similar retrieval is inferior to selecting the most-similar one, we also observe that logit-based retrieval is more effective than feature-based one. We assume this is due to the image classification task is more related to logit value rather than feature similarity.

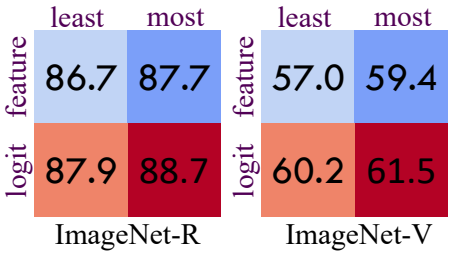


FIGURE 5.8. Performance analysis on different retrieval strategies.

Effect of In-Context Exemplars with Distribution Shift. When the support set suffers from a distribution shift from the target OOD dataset, whether DICL can still perform robustly remains to be validated. Hence, we leave one domain out as our support set and leverage the rest domains as our target OOD dataset. In comparison, we choose a small hold-out data split as the support set which shares the same distribution as the OOD dataset. The results are shown in Figure 5.9 left. Surprisingly, we find that the performance is not influenced by the distribution shift, which demonstrates that our MVT can still be effective when exemplars are retrieved from different distributions.

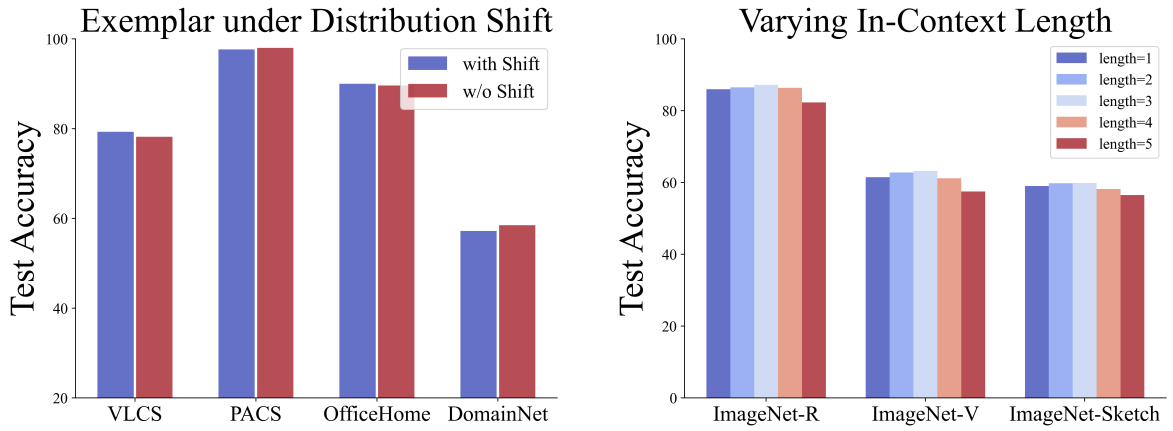


FIGURE 5.9. Performance analysis of in-context exemplars: (left) under distribution shift; (right) varying the exemplar length.

Effect of Varying In-Context Length. Further, we analyze the effect of increasing exemplar length during inference. Particularly, we consider one positive and negative exemplar pair as length 1. Here we vary the length from 1 to 5 and show the results in Figure 5.9 right. We observe slight improvement when the length gradually increases which is consistent with the theoretical findings [298]. However, when the length is longer than 4 the performance drops and the predictions of MLLM become unstable which could be other than "True" or "False". This might be due to the limited capacity of MLLMs on handling a certain amount of information, which is worth conducting studies on sophisticated MLLMs in the future.

Performance of OOD Detection. At last, we consider a more challenging scenario where data from open classes could exist in the target dataset. Here we simulate this situation by choosing 60% of the classes as closed classes and the rest are open classes. To detect such open-class data, i.e., OOD detection [28, 307][6], we use the vision model prediction confidence as a baseline and compare it with the MLLM diagnosing confidence as well as the detection score Δ in Equation (5.4). The result is shown in Figure 5.10. In the upper row, we observe the similar clearly distinguishable distributions using our score Δ as in Figure 5.6. In the lower row, we show the F1 score of each detection criterion under a threshold varied from 0 to 1 on three datasets. When a criterion produces confidence larger than the threshold, it would predict as close-class data, other as open-class ones. Based on the result, we find

[6]Note that OOD detection here is different from the previous setting: here we focus on detecting open-class data, and previous one focuses on detecting prediction errors.

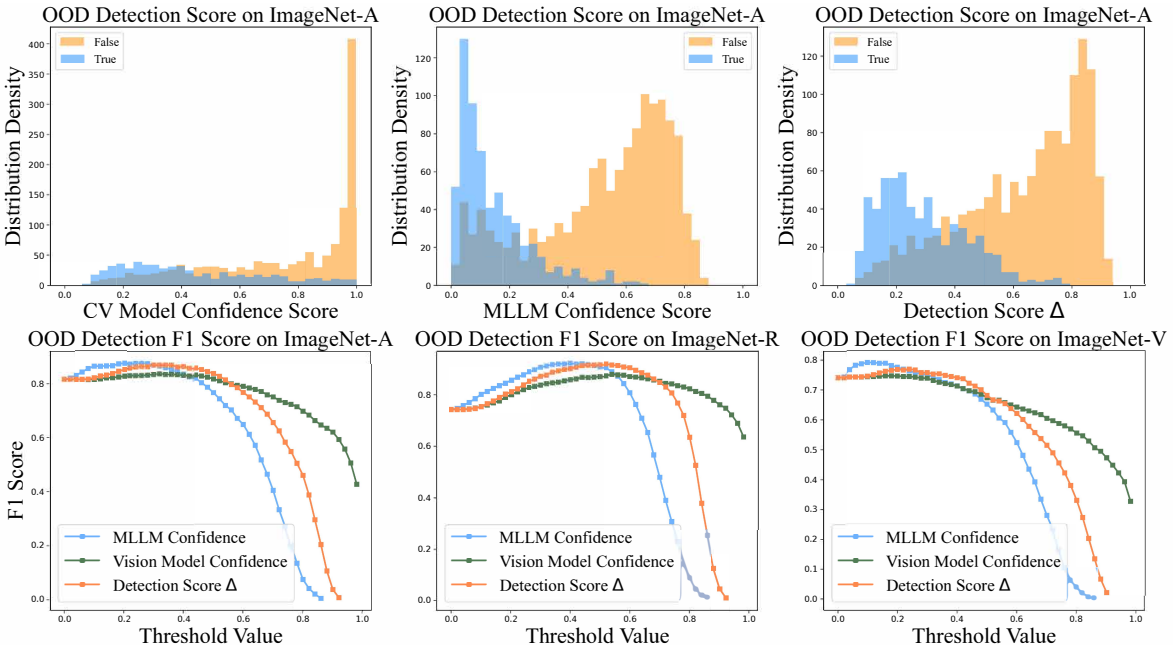


FIGURE 5.10. OOD detection analysis. Upper: Detection score distribution on ImageNet-A; Lower: F1 scores of vision model confidence, MLLM diagnosing confidence, and our Δ score in ImageNet-A, ImageNet-R, and ImageNet-V.

that MLLM achieves better detection performance when the threshold is small, but vision model confidence is relatively better when the threshold is large, i.e., MLLM can effectively detect open classes while vision models are better at recognizing close classes. However, an effective detection should have a reasonable threshold value that is neither too large nor too small and meanwhile has a high F1 score. Hence, by combining them together, our detection score Δ can achieve the best F1 score when the threshold is around the middle range.

5.5 Conclusion

In this Chapter, we propose a novel paradigm of fine-tuning vision models via leveraging MLLMs to improve visual robustness on downstream OOD tasks. Specifically, we effectively estimate a transition matrix to help find the most probable noisy classes. By using a positive exemplar and a negative exemplar retrieved based on the noisy classes, we can

conduct DICL to rectify incorrect vision model predictions through two stages dubbed diagnosing and therapy. Thanks to the rectified predictions, the robustness of vision models can be further improved through fine-tuning. We conduct detailed theoretical analysis and extensive quantitative and qualitative experiments to justify the proposed method. Our framework can significantly reduce the cost of training vision models and provide insights into many visual recognition problems, such as OOD detection, OOD generalization, weakly-supervised learning, etc.

CHAPTER 6

Out-of-Modal Generalization

The world is understood from various modalities, such as appearance, sound, and language. Since each modality only partially represents objects in a certain meaning, leveraging additional ones is beneficial in both theory and practice. However, exploiting novel modalities normally requires cross-modal pairs corresponding to the same instance, which is extremely resource-consuming and sometimes even impossible, making knowledge exploration of novel modalities largely restricted. To seek practical multi-modal learning, here we study Out-of-Modal (OOM) Generalization as an initial attempt to generalize to an unknown modality without given instance-level modal correspondence. Specifically, we consider Semi-Supervised and Unsupervised scenarios of OOM Generalization, where the first has scarce correspondences and the second has none, and propose Connect&Explore (COX) to solve these problems. COX first connects OOM data and known In-Modal (IM) data through a variational information bottleneck framework to extract shared information. Then, COX leverages the shared knowledge to create emergent correspondences, which is theoretically justified from an information-theoretic perspective. As a result, the label information on OOM data emerges along with the correspondences, which helps explore the OOM data with unknown knowledge, thus benefiting generalization results. We carefully evaluate the proposed COX method under various OOM generalization scenarios, verifying its effectiveness and extensibility.

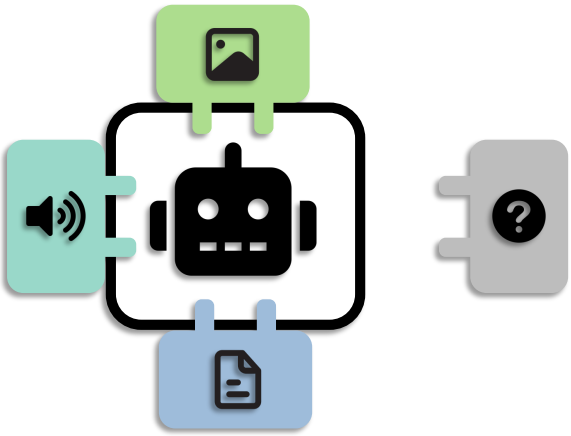


FIGURE 6.1. AI is enhanced as more modalities are incorporated, so how can AI learns from novel modalities based on the ones it already know?

6.1 Introduction

To understand the world, we use various data *modalities*, such as image [3, 308, 309] and text [310, 5]. Each modality describes objects through a certain physical perspective, thus contributing to understanding objects. Therefore, *multi-modal learning* (MML) [246, 311, 312, 226, 313, 314, 315] which learns from multiple modality data has been a core research topic in AI. Thanks to the utilization of various modalities, the learning performance has shown benefits on various tasks compared to uni-modal learning [316, 317, 226, 318], such as cross-modal retrieval [280, 319, 320], human-computer interaction [321, 322], and robotics [323, 324].

However, existing states of the art are not satisfactory, and emerging modalities need to be leveraged effectively just like the relatively new data modalities of geomagnetic fields [325], sound waves [326], and electromagnetic waves [327]. Therefore, emerging technologies have constantly leveraged new sensors to enhance their performance. For example, Embodied AIs [328] already possess abilities like 3D vision and language, but they are still exploring novel skills, such as tactile and bio-sensing. Since it is hard to leverage such uncommon and inexperienced skills in practice, adapting the knowledge from common modalities to understand the novel ones could be beneficial, as shown in Figure 6.1. In practice,

most existing MML investigations [226, 295, 329, 330] require *instance-level modal correspondence*, i.e., multi-modal data are paired with the same instance, which is often hard to satisfy in real-world scenarios when facing novel modalities [331, 332, 318, 333]. For a robotic example, some modalities are common and easy to acquire, *e.g.*, vision and language. However, others like tactility need special sensors to resample from the same objects seen or spoken. Unfortunately, the resample could no longer be accessible in practice. As a result, the new modalities usually have incomplete or even no correspondence, which could seriously block the knowledge interaction across modalities and hinder the benefits brought by MML. Hence, a question naturally occurs: *Do we really need instance-level modal correspondence to explore novel modalities?*

This Chapter studies a practical yet unexplored problem named *Out-of-Modal (OOM) Generalization*. Particularly, given several modalities, *i.e.,* In-Modal (IM) data, the goal is to generalize to an unknown modality without or sometimes only with scarce correspondence. Such a setting implies the real-world utilization of novel modalities: Even though human perception is limited to certain modalities, *e.g.*, touch, sight, sound, and smell, we can still understand unperceivable ones such as magnetism by utilizing inherently-possessed senses, *e.g.*, feel the force when pulling two magnets together; or see the magnetic field by observing the alignment of iron filings around a magnet.

Based on this insight, we utilize IM perceptors that contain prior knowledge to encode known IM data, which can be implemented using existing MML models [226, 295, 330, 329], and an OOM learner which learns novel modalities without any prior knowledge. By analyzing the interactions between latent features, we show theoretically and empirically that the OOM learner can be trained to gradually discover the OOM knowledge, as shown in Figure 6.2. First, we consider *semi-supervised OOM generalization* where few correspondences are given. Based on the correspondence, we can capture the prior probability distribution and learn mappings that connect OOM data and IM data. Through an information-theoretic perspective, we propose *Connect&Explore* (COX), which encourages the agreement on mappings across modalities, further sharing the cross-modal knowledge and exploring OOM knowledge. Then, we extend COX to an *unsupervised OOM generalization* scenario where

there is no instance-level correspondence at all. To tackle such a challenge, we enhance the OOM-IM connections by maximizing cross-modal interaction. First, we select data pairs from cross-modal mappings according to feature similarity. By assuming that the data pairs closing to OOM mappings can be considered as correspondence, we can create emerging correspondence and solve the unsupervised case via the semi-supervised solution. To validate the proposed COX, we carefully design experiments using various multi-modal datasets to validate its effectiveness. Moreover, we provide extensive analyses to understand the OOM problem and inspire future research. To sum up, our contributions are threefold:

- We discover a novel and practical problem named OOM Generalization, which aims to explore a novel modality using the knowledge from known modalities.
- We consider two typical situations: Semi-Supervised OOM generalization and Unsupervised OOM generalization, and propose a Connect&Explore framework to tackle both problems from an information-theoretic perspective.
- We conduct extensive experiments to tackle the OOM generalization on various datasets and provide intuitive insights to help inspire future research.

6.2 Related Work

Modality Generalization [334] generally focuses on leveraging the knowledge from some modalities and generalizing to another one. Existing studies are conducted in different settings and with various tasks. Cross-modal fine-tuning mimics transfer learning by adapting the distribution of IM data to OOM data using the same model. [335] proposed to conduct distribution alignment to achieve this goal which requires both pre-trained knowledge and labeled target modality data. Based on a similar problem setting, [336] designed a gradual modality generation scheme that selects the top-k active feature patches from target modalities, and replaces them with source modalities patches. Such a progressive strategy can align target modal data to ensure generalization. Cross-Modal Generalization uses separate encoders and focus on generalizing to a different modality data from the same instance.

TABLE 6.1. A comparison of different MML problems and their corresponding settings.

Problem	References	IM Knowledge	OOM Knowledge	Correspondence
Cross-Modal Fine-Tuning	[335, 336]	pre-trained & labeled	labeled	✗
Cross-Modal Generalization	[332]	pre-trained & labeled	pre-trained	✔
	[333]	pre-trained & labeled	pre-trained & labeled	✔
MML w/o labeled MM Data	[331]	partially labeled	partially labels	✔
OOM Generalization	Section 6.3.3	pre-trained & labeled	scarcely labeled	A few
	Section 6.3.4	pre-trained & labeled	✗	✗

[332] used meta-learning to align OOM data to IM space and generalize to OOM tasks dynamically. [333] studied a different setting where IM and OOM data are both known during training. Then, a unified representation space is learned to help with the downstream generalization of OOM data. Some other studies consider generalization when all modalities are available, [337] studied cross-modal generalization without paired data, [338] applied the information bottleneck to CLIP training, [339] conducted multi-modal fusion under limited clinical data, and [340] considered domain generalization with fully-paired multi-modal data. A recent study MML without Labeled Multi-Modal Data [331] proposed a different setting where both IM and OOM data have labels, but they are not paired. Instead, additional unlabeled paired multi-modal data is given for learning the interaction between modalities. Moreover, [341] understood the interactions and applied it to knowledge distillation. Except for cross-modal fine-tuning which follows transfer learning, existing MML works mostly require instance-level correspondence. This work proposes OOM Generalization, where there is no correspondence and the OOM knowledge is barely provided. The comparison of related works is shown in Table 6.1.

Modality Binding aims to learn a joint embedding space across different modalities. Contrastive Language-Image Pre-training CLIP [226] is the first work that aligns image with language data. Then, ImageBind [295] proposed to use vision modalities to bind various modalities into the same representation space. Further, LanguageBind [330] proposed using language as an alternative solution, which binds various modalities similarly. Recently, FreeBind [329] extended the existing unified space into an additional expert space. Specifically,

two types of binding were considered, namely space displacement bond and space combination bind. Since modality binding often requires a large amount of data with correspondence, the selected modalities are often quite common. Therefore, the OOM generalization problem can take advantage of the development of modality binding by leveraging the encoders as our IM perceptors to learn novel modalities.

6.3 OOM Generalization

In this section, we first formalize the OOM generalization setting. Then, we demonstrate the proposed method. Further, we consider a Semi-Supervised case where a few correspondences are available and an Unsupervised scenario where there is no correspondence, showing that the proposed method can successfully tackle both settings and effectively leverage unpaired OOM data.

6.3.1 Problem Setting

In OOM generalization, we are given a set of known modalities $\{\mathcal{M}_1^{\mathrm{I}}, \ldots, \mathcal{M}_K^{\mathrm{I}}\}$ where $\mathcal{M}_{k\in\{1,\ldots,K\}}^{\mathrm{I}} = \{(x_{k,i}^{\mathrm{I}}, y_{k,i}^{\mathrm{I}})_{i=1}^{N} \in \mathcal{X} \times \mathcal{Y}\}$ is composed of N number of labeled IM examples with its subscript i denoting the correspondence across different modalities. Moreover, we have an unknown modality $\mathcal{M}^{\mathrm{O}} = \{(x_j^{\mathrm{O}})_{j=1}^{M}\}$ containing M unlabeled OOM examples. In some cases, it is possible to obtain few correspondences with IM data, then our OOM data could be $\mathcal{M}^{\mathrm{O}} = \{(x_i^{\mathrm{O}}, y_i^{\mathrm{O}})\}_{i=1}^{L} \cup \{(x_j^{\mathrm{O}})\}_{j=L+1}^{M}$, where $L \ll M$ and the subscript i traces the corresponding IM data instance and label.

To tackle OOM generalization, we propose a learning framework as shown in Figure 6.2. Particularly, we use a set of IM perceptors $\{g_1^{\mathrm{I}}, \ldots, g_K^{\mathrm{I}}\}$ to perceive IM data, which can be realized by many existing modality-binding models, such as ImageBind [295] and LanguageBind [330]. Then, the features of IM data are obtained via $z_k^{\mathrm{I}} = g_k^{\mathrm{I}}(x_k^{\mathrm{I}})$. Moreover, we use an OOM learner g^{O} to learn features z^{O} from OOM data through $z^{\mathrm{O}} = g^{\mathrm{O}}(x^{\mathrm{O}})$. Our goal is to effectively generalize to OOM data by exploring the relationships between the OOM feature z^{O} and IM features $\{z_k^{\mathrm{I}}\}_{k=1}^{K}$. Note that we only focus on the generalization performance of

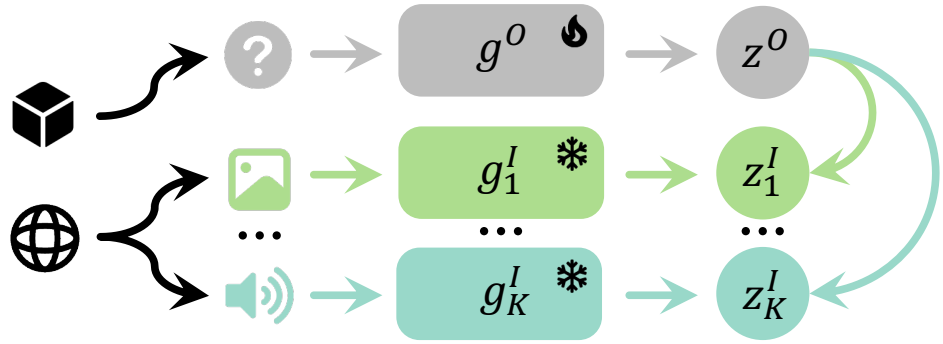


FIGURE 6.2. Learning framework of our OOM generalization.

OOM data, the improvement of learning IM data is not the goal of this Chapter. Therefore, we freeze the parameters of all IM perceptors and only train the OOM learner during experiments. On top of the above models, we further define classifiers $h^{\mathrm{O}}(x^{\mathrm{O}}) := h^{\mathrm{O}}(x^{\mathrm{O}}; g^{\mathrm{O}})$ and $h_k(x_k^{\mathrm{I}}) := h_k(x_k^{\mathrm{I}}; g_k^{\mathrm{I}})$ that make predictions.

6.3.2 Methodology: Connect&Explore (COX)

Here we elucidate the proposed method based on the interactive relationship between modalities [331, 342]. Specifically, the total information of two modalities under a certain task is decomposed into (1) *commonality*[1] which indicates common attributes across modalities, (2) *uniqueness* that is only presented in each modality, and (3) *synergy* denoting the emerging information when modalities are presented together. Note that we do not consider (3) in this Chapter as our goal is generalizing to OOM data.

To generalize to an unknown modality based on common ones, we aim to extract the commonality that can help partially comprehend OOM data based on IM data. Then, we model the posterior distribution of OOM data by selecting anchor points with minimum uniqueness. To this end, the OOM generalization can be successfully established. The proposed COX method comprises two steps: (1) learning connections by mapping IM data to OOM data to extract commonality, and (2) exploring high uniqueness OOM data by matching their posterior to high-commonality OOM data.

[1]It is originally termed "redundancy" which is negative. However, such property is quite positive for tackling our problem, and hence we rename it "commonality".

Connection through Commonality aims to capture common knowledge across modalities using generative models [317]. Here we follow the variational information bottleneck (VIB) framework [343] to achieve this goal. We assume that given IM data X^{I} and OOM data X^{O}, the latent variable V extracted from X^{I}[2], and label Y, the joint distribution is factorized as

$$p(X^{\mathrm{I}}, X^{\mathrm{O}}, V, Y) = p(V, Y|X^{\mathrm{O}}, X^{\mathrm{I}})p(X^{\mathrm{O}}|X^{\mathrm{I}})P(X^{\mathrm{I}}), \tag{6.1}$$

where we assume $p(V, Y|X^{\mathrm{O}}, X^{\mathrm{I}}) = p(V|X^{\mathrm{I}})p(Y|X^{\mathrm{I}})$, corresponding to the Markov chains $V \leftrightarrow X^{\mathrm{I}} \leftrightarrow X^{\mathrm{O}}$ and $X^{\mathrm{I}} \leftrightarrow Y \not\leftrightarrow X^{\mathrm{O}}$. Such an assumption means that V is not related to X^{O} [343] and the given label Y is not directly connected to X^{O} under our OOM setting. Intuitively, given an IM datum, i.e., dog image, it is sufficient to infer the label "dog", and the same for inferring from an unknown OOM datum, i.e., dog bark. Thus, in common multi-modal settings, the label prediction using IM information dog image is not further conditioned on OOM knowledge dog bark, because here the OOM knowledge is redundant when IM data is given.

Our goal is to extract valuable knowledge from IM data to leverage OOM data by maximizing the information commonality [331, 342]:

$$\max I(X^{\mathrm{O}}; X^{\mathrm{I}}; Y) = I(X^{\mathrm{O}}; X^{\mathrm{I}}) - I(X^{\mathrm{O}}; X^{\mathrm{I}}|Y), \tag{6.2}$$

where $I(X^{\mathrm{O}}; X^{\mathrm{I}}; Y)$ denotes the mutual information between X^{O} and X^{I} regarding the task Y, i.e., the label; and $I(X^{\mathrm{O}}; X^{\mathrm{I}}|Y)$ indicates the conditional mutual information irrelevant to Y. We start with the first term:

$$I(X^{\mathrm{O}}; X^{\mathrm{I}}) = \int dx^{\mathrm{O}} dx^{\mathrm{I}} p(x^{\mathrm{O}}, x^{\mathrm{I}}) \log \frac{p(x^{\mathrm{O}}, x^{\mathrm{I}})}{p(x^{\mathrm{O}})p(x^{\mathrm{I}})} = \int dx^{\mathrm{O}} dx^{\mathrm{I}} p(x^{\mathrm{O}}, x^{\mathrm{I}}) \log \frac{p(x^{\mathrm{O}}|x^{\mathrm{I}})}{p(x^{\mathrm{O}})}, \tag{6.3}$$

where $p(x^{\mathrm{O}}|x^{\mathrm{I}}) = \int dv p(x^{\mathrm{O}}, v|x^{\mathrm{I}}) = \int dv p(x^{\mathrm{O}}|v)p(v|x^{\mathrm{I}})$ can be approximated via a decoder $q(x^{\mathrm{O}}|v)$. Since the Kullback Leibler (KL) divergence is always non-negative, we have $\mathrm{KL}[p(X^{\mathrm{O}}|V) \parallel q(X^{\mathrm{O}}|V)] \geq 0 \Rightarrow \int dx^{\mathrm{O}} p(x^{\mathrm{O}}|v) \log p(x^{\mathrm{O}}|v) \geq \int dx^{\mathrm{O}} p(x^{\mathrm{O}}|v) \log q(x^{\mathrm{O}}|v)$,

[2]Note that the latent variable V here is different from the feature representation z^{I} and z^{O}.

and leveraging Jensen's inequity, we can have

$$I(X^{\mathrm{O}};X^{\mathrm{I}}) \geq \int dx^{\mathrm{O}}dx^{\mathrm{I}}p(x^{\mathrm{O}},x^{\mathrm{I}})\log\frac{\int dvq(x^{\mathrm{O}}|v)p(v|x^{\mathrm{I}})}{p(x^{\mathrm{O}})} \tag{6.4}$$

$$\geq \int dx^{\mathrm{O}}dx^{\mathrm{I}}dvp(x^{\mathrm{O}},x^{\mathrm{I}})\log q(x^{\mathrm{O}}|v)p(v|x^{\mathrm{I}}) + H(X^{\mathrm{O}}), \tag{6.5}$$

where the last term is independent of our optimization process. Further, we rewrite:

$$p(x^{\mathrm{O}},x^{\mathrm{I}}) = \int dvp(x^{\mathrm{O}},x^{\mathrm{I}},v) = \int dvp(x^{\mathrm{I}})p(x^{\mathrm{O}}|x^{\mathrm{I}})p(v|x^{\mathrm{I}}). \tag{6.6}$$

Then, we have the following lower bound:

$$I(X^{\mathrm{O}};X^{\mathrm{I}}) \geq \int dx^{\mathrm{O}}dx^{\mathrm{I}}dvp(x^{\mathrm{I}})p(x^{\mathrm{O}}|x^{\mathrm{I}})p(v|x^{\mathrm{I}})\log q(x^{\mathrm{O}}|v)p(v|x^{\mathrm{I}}), \tag{6.7}$$

which is realized by sampling from the joint data distribution, the latent variable from our encoder $p(v|x^{\mathrm{I}})$, and the tractable variational approximation $q(x^{\mathrm{O}}|v)$.

Similarly, we can upper-bound the second term $I(X^{\mathrm{O}};X^{\mathrm{I}}|Y)$:

$$I(X^{\mathrm{O}};X^{\mathrm{I}}|Y) \leq \int dx^{\mathrm{O}}dx^{\mathrm{I}}dyp(x^{\mathrm{O}},x^{\mathrm{I}},y)\log p(y|x^{\mathrm{I}})p(x^{\mathrm{O}}|x^{\mathrm{I}})p(x^{\mathrm{I}}) - \log h^{\mathrm{O}}(y|x^{\mathrm{O}}), \tag{6.8}$$

where $h^{\mathrm{O}}(y|x^{\mathrm{O}})$ is our classifier model for predicting OOM data. To this end, we can lower-bound our objective by combining Equations. (6.7) and (6.8):

$$\begin{aligned} I(X^{\mathrm{O}};X^{\mathrm{I}};Y) \geq & \int dx^{\mathrm{O}}dx^{\mathrm{I}}dvp(x^{\mathrm{I}})p(x^{\mathrm{O}}|x^{\mathrm{I}})p(v|x^{\mathrm{I}})\log q(x^{\mathrm{O}}|v)p(v|x^{\mathrm{I}}) \\ - & \int dx^{\mathrm{O}}dx^{\mathrm{I}}dyp(x^{\mathrm{O}},x^{\mathrm{I}},y)\log p(y|x^{\mathrm{I}})p(x^{\mathrm{O}}|x^{\mathrm{I}})p(x^{\mathrm{I}}) + \log h^{\mathrm{O}}(y|x^{\mathrm{O}}) = \mathcal{L}_{\mathrm{con}}. \end{aligned} \tag{6.9}$$

The above lower bound contains two parts: (1) OOM data reconstruction where we reconstruct X^{O} using the latent V and (2) OOM data label prediction where we model the label distribution Y. In practice, we can approximate $p(x^{\mathrm{O}},x^{\mathrm{I}},y)$ using empirical samples from IM and OOM data. Moreover, we use encoder $p(v|x^{\mathrm{I}})$ without any prior assumptions because we can leverage the feature distribution from the pre-trained IM perceptors. Additionally, a classifier $h(y|x^{\mathrm{O}})$ is optimized to categorize OOM data based on given labels. Empirically,

we can minimize

$$\mathcal{L}_{\text{con}} := \frac{1}{M}\sum_{i=1}^{M} \|x_i^{\text{O}} - q(x_i^{\text{O}}|v_i)p(v_i|x_i^{\text{I}})\|_2^2 - \log h^{\text{O}}(y_i|x_i^{\text{O}}), \tag{6.10}$$

where we use the reconstruction error $\|\cdot\|_2^2$[3] to realize the log-likelihood $q(x^{\text{O}}|v)p(v|x^{\text{I}})$, as similarly done by Kingma and Welling [83]. After building the connections, we can ensure the task-relevant information shared across modalities is learned, which helps partially understand OOM data regarding its commonality. However, note that the second term in Equation (6.9) is not fully leveraged which contains $p(y|x^{\text{I}})$ modeled by the IM perceptors. Take a step further, we can obtain $-\int dx^{\text{O}}dx^{\text{I}}dy p(x^{\text{O}}, x^{\text{I}}, y) \log \frac{p(y|x^{\text{I}})p(x^{\text{O}}|x^{\text{I}})p(x^{\text{I}})}{h^{\text{O}}(y|x^{\text{O}})}$. Since $p(x^{\text{O}}|x^{\text{I}})p(x^{\text{I}})$ is fixed in label prediction, we can derive $-\text{KL}(p(y|x^{\text{I}}) \parallel h^{\text{O}}(y|x^{\text{O}}))$ which implies that the label information related IM data can be harnessed to explore commonality. Next, we demonstrate how the commonality helps OOM generalization, and provide a solution to explore uniqueness.

Exploration of Uniqueness can be achieved via selecting and exploring the OOM data with high uniqueness. To identify these data, we can leverage the agreement and disagreement achieved by the optimal classifiers from various IM data. Our final goal is to optimize via

$$\min_{h^{\text{O}}} \text{KL}(h^{\text{O}}(y|x_d^{\text{O}}) \parallel h^{\text{O}}(y|x_a^{\text{O}})), \text{where } x_d^{\text{O}} \in \mathcal{D}, x_a^{\text{O}} \in \mathcal{A}, \tag{6.11}$$

in which h_1^* and h_2^* denote the optimal classifiers found in two IM data x_1^{I} and x_2^{I}, respectively, and x_d^{O} and x_a^{O} are selected from OOM data with modality disagreement $\mathcal{D} := \{x^{\text{O}} : h_1^*(x^{\text{O}}) \neq h_2^*(x^{\text{O}})\}$ and agreement $\mathcal{A} := \{x^{\text{O}} : h_1^*(x^{\text{O}}) = h_2^*(x^{\text{O}})\}$, respectively. Here we use two in-modalities for simplicity, but the conclusion can be extended to multiple modalities. Moreover, the data with agreement is considered anchor points that guide the exploration of those with disagreement. This objective aims to match the posterior of OOM data with uniqueness $h^{\text{O}}(y|x_d^{\text{O}})$ to the one of anchor points $h^{\text{O}}(y|x_a^{\text{O}})$. To justify this, we first define modality disagreement:

[3]Though training generative models in input space is computationally inefficient, we propose to connect modalities in the feature space in experiments. Therefore, the raw data x is replaced by latent feature z.

DEFINITION 6 (Modality disagreement). *Given X_1, X_2 and target Y, as well as their corresponding optimal classifiers h_1^* and h_2^*, their modality disagreement is defined as $\alpha(h_1^*, h_2^*) = \mathbb{E}_{p(x_1,x_2)}[d(h_1^*, h_2^*)]$ where $d : \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}^+$ is a distance function in the label space scoring the disagreement between h_1^* and h_2^*.*

THEOREM 7. *Given two Bayes' optimal classifiers h_1^* and h_2^* from two in-modalities, under relaxed triangle inequality, inverse Lipschitz condition, and classifier optimality assumptions [344], the modalities disagreement is upper-bounded by*

$$\alpha(h_1^*, h_2^*) \leq I(X^{\mathrm{O}}, X_2^{\mathrm{I}}, Y|X_1^{\mathrm{I}}) + I(X^{\mathrm{O}}, X_1^{\mathrm{I}}, Y|X_2^{\mathrm{I}}) + 2I(X^{\mathrm{O}}, Y|X_1^{\mathrm{I}}, X_2^{\mathrm{I}}). \tag{6.12}$$

Finally, based on the decomposition of the task-related mutual information X^{O}:

$$I(X^{\mathrm{O}}, Y) = I(X^{\mathrm{O}}, X_2^{\mathrm{I}}, Y|X_1^{\mathrm{I}}) + I(X^{\mathrm{O}}, X_1^{\mathrm{I}}, Y|X_2^{\mathrm{I}}) + I(X^{\mathrm{O}}, Y|X_1^{\mathrm{I}}, X_2^{\mathrm{I}}) + I(X^{\mathrm{O}}, X_1^{\mathrm{I}}, X_2^{\mathrm{I}}, Y), \tag{6.13}$$

as shown in Figure 6.3, we can achieve

$$\alpha(h_1^*, h_2^*) \leq I(X^{\mathrm{O}}, Y) - I(X^{\mathrm{O}}, X_1^{\mathrm{I}}, X_2^{\mathrm{I}}, Y) + I(X^{\mathrm{O}}, Y|X_1^{\mathrm{I}}, X_2^{\mathrm{I}}), \tag{6.14}$$

where the first term denotes the overall information, the second term indicates the commonality shared between all modalities, and the third term stands for the uniqueness only preserved in OOM data. Intuitively, when we try to increase the modality disagreement, the commonality is decreased and OOM uniqueness is increased, which successfully justifies our learning objective: In order to explore the uniqueness of OOM data, we can explore the ones with high modality disagreement; conversely, the OOM

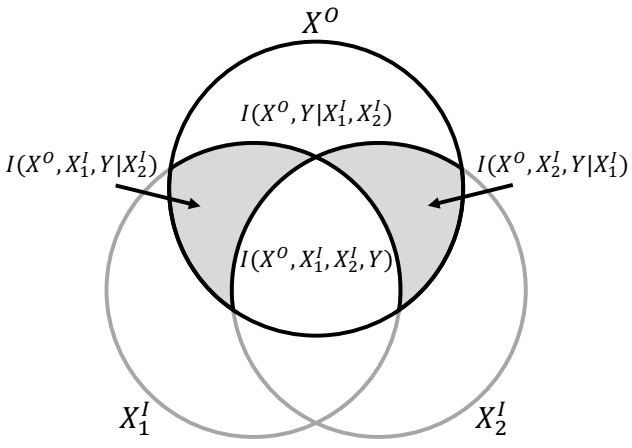


FIGURE 6.3. Decomposition of $I(X^{\mathrm{O}}, Y)$.

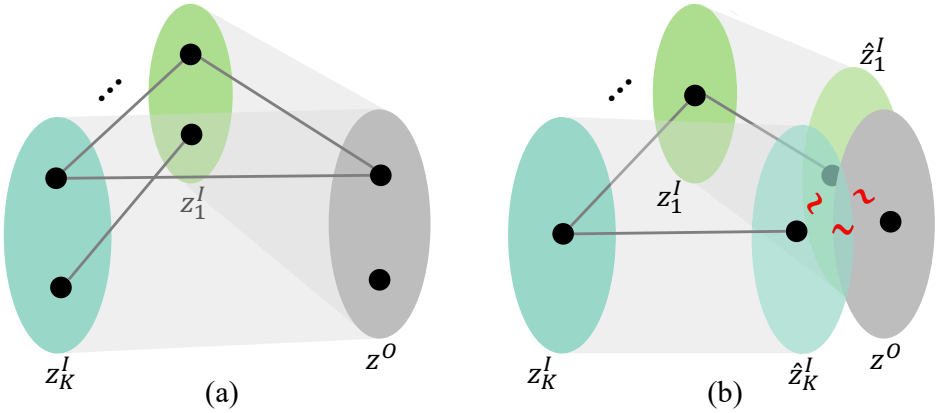


FIGURE 6.4. Two scenarios: (a) Semi-Supervised OOM Generalization and (b) Unsupervised OOM Generlaizaiton.

data with high commonality and low uniqueness is found where agreement is achieved among h_1^* and h_2^*. Therefore, we select such data as anchor points that provide informative guidance to help explore uniqueness.

Next, we consider two realistic scenarios of OOM generalization and demonstrate how the proposed COX method can tackle them.

6.3.3 Semi-Supervised OOM Generalization

We start with a semi-supervised case where a few correspondences are available in OOM data, as shown in Figure 6.4 (a). Based on the VIB framework proposed in Section 6.3.2, we first leverage the OOM data $\{(x_i^{\mathrm{O}}, y_i^{\mathrm{O}})\}_{i=1}^L$ corresponding to IM data $\{(x_{k,i}^{\mathrm{I}}, y_{k,i}^{\mathrm{O}})\}_{i=1}^L, \forall k \in \{1, \ldots, K\}$ to build K connections using additional generative models that can be trained via a point-to-point mapping. As a result, the mappings on the OOM feature space can successfully match the OOM feature distribution, which allows us to directly apply IM data posteriors to select and explore the uniqueness of OOM data. Hence, we formulate our objective as

$$\min_{h^{\mathrm{O}}} \mathcal{L}_{\mathrm{ssl}} := \frac{1}{L}\sum_{i=1}^{L} \mathrm{CE}(h^{\mathrm{O}}(x_i^{\mathrm{O}}), y_i^{\mathrm{O}}) + \frac{1}{L+|\mathcal{D}|}\sum_{x_{d,j}\in\mathcal{D}}\sum_{x_{i=1}^L} \mathrm{KL}(h^{\mathrm{O}}(x_{d,j}^{\mathrm{O}}) \,\|\, h^{\mathrm{O}}(x_i^{\mathrm{O}}); h_1^*, h_2^*), \tag{6.15}$$

where the first term exploits labeled OOM data with correspondence and the second term explores OOM data $\mathcal{D}$ with modality disagreement by minimizing its KL divergence from the label posterior. Through the above objective, we can maximally exploit the uniqueness of OOM data to achieve effective OOM generalization.

6.3.4 Unsupervised OOM Generalization

As for the unsupervised case, we propose two-phase training: (1) we first conduct a warm-up training to initialize the OOM feature space and the connection, and (2) then, we enhance the connection by creating emergent correspondence and further exploring OOM data.

Specifically, we select anchor points from OOM data by directly applying modality agreement among all Bayes' optimal classifiers from IM data via

$$\mathcal{A}_{\text{sorted}} = \text{SORT}_T(\mathcal{A}, \frac{1}{K}\sum_{k=1}^{K} \max h_k^*(x^{\text{O}})), \text{where } \mathcal{A} = \{\forall x^{\text{O}} \in \mathcal{M}^{\text{O}}\colon h_1^*(x^{\text{O}}) = \cdots = h_K^*(x^{\text{O}})\}, \tag{6.16}$$

where the $\text{SORT}_T(\cdot,\cdot)$ is a sort function, which ranks each element x^{O} in $\mathcal{A}$ based on the value of $\frac{1}{K}\sum_{k=1}^{K} \max h_k^*(x^{\text{O}})$ from large to small. Here, we select anchor points with the top-T largest likelihood averaged over all K IM classifiers. Then, we warm up the OOM learner via minimizing cross-entropy loss $\min \frac{1}{T}\sum_{x^{\text{O}} \in \mathcal{A}_{\text{sorted}}} \text{CE}(h^{\text{O}}(x^{\text{O}}), \arg\max h_k^*(x^{\text{O}}))$. Additionally, we also warm up the connection by leveraging class-wise information. Specifically, we compute the cluster centroids for each modality via $\frac{1}{|\mathcal{C}_y|}\sum_{x_i^{\text{O}} \in \mathcal{C}_y := \{x^{\text{O}}: h^{\text{O}}(x^{\text{O}}) = y, y \in \mathcal{Y}\}} z_i^{\text{O}}$ and pair them to each IM centroid correspondingly. To this end, we can build up initial connections by following the VIB framework.

After the warm-up, we aim to further enhance both our connection and OOM exploration by creating emergent correspondence, as shown in Figure 6.4 (b). To tackle this, we map all IM data into the OOM feature space. If an OOM feature is close to all mappings $v_{k,i}, \forall k = \{1, \ldots, K\}$, then they can form a strong correspondence. Further, we select such OOM data as anchor points, which is further labeled the same as the corresponding IM data. Formally,

we optimize OOM learners via

$$\min_{h^{\mathrm{O}}} \mathcal{L}_{\mathrm{uns}} := \frac{1}{|\mathcal{A}|} \sum_{(x_a^{\mathrm{O}}, y) \in \mathcal{A}} \mathrm{CE}(h^{\mathrm{O}}(x_a^{\mathrm{O}}), y) + \frac{1}{|\mathcal{A}|+|\mathcal{D}|} \sum_{x_d^{\mathrm{O}} \in \mathcal{D}} \sum_{x_a^{\mathrm{O}} \in \mathcal{A}} \mathrm{KL}(h^{\mathrm{O}}(x_d^{\mathrm{O}}) \,\|\, h^{\mathrm{O}}(x_a^{\mathrm{O}}); h_1^*, h_2^*), \tag{6.17}$$

where $\mathcal{A}$ denotes the updated anchor points which are realized by sorting the Euclidean distance: $\mathcal{A} := \mathrm{SORT}_S(\{(x_j^{\mathrm{O}}, y_i^{\mathrm{I}})\}_{j=1}^M, -\min_{i \in \{1,\ldots,N\}} \frac{1}{K} \sum_{k=1}^K \|z_j^{\mathrm{O}} - v_{k,i}\|)$, where the first term computes the cross-entropy loss from the anchor points, and the second term calculates the KL divergence between the OOM data with modality disagreement and the anchor points.

After these two steps, we can effectively tackle the unsupervised OOM generalization. In practice, we connect modalities and select anchor points in the feature space, and hence our application to both two scenarios can be efficient. In the next section, we carefully conduct extensive experiments to justify the effectiveness and extendibility of the proposed COX method under various settings.

6.4 Experiment

In our experiments, we first elucidate the experimental details. Then, we provide performance comparisons to various baseline methods on different datasets. Finally, we conduct empirical analyses to provide an intuitive understanding of the proposed method.

6.4.1 Implementation Details

Datasets. We consider datasets with at least three modalities: (1) TVL dataset [345] contains tactile sensing, RGB image, and class name which can be transformed into language; (2) LLVIP [346] dataset has infrared thermal data, RGB image, and annotations for pedestrian detection. We follow [330] to crop the pedestrian and background which stand for two classes. Further, we use the OpenAI template [226] to create language description; (3) NYU-D dataset [347] contains RGB image, depth data, and class name that can be transformed into language description as well; (4) VGGS dataset [348] includes video data, corresponding sound, and the language description; (5) MSR-VTT [349] includes videos and

text description, we break down the videos into video frames and the audio data; (6) MOSEI dataset [350] contains videos from 7 classes of emotions, we extract audio data from the videos and use the emotion type to create language descriptions.

Models. We employ two types of IM perceptors, namely ImageBind [295] and LanguageBind [330] which correspondingly contain 6 and 5 encoders to process different modalities. We select one modality for each experiment as OOM and then choose the rest as IM. For IM data, we use the existing encoders to extract their features. As for OOM data, we conduct preprocessing to ensure its compatibility. Then, we initialize an OOM learner from scratch using ViT-T/16 to learn from the OOM data using the guidance from IM perceptors. Note that for the TVL dataset, there are no existing encoders to process tactile modality. Therefore, when the tactile modality is chosen as IM data, we fine-tune the encoder using contrastive learning on the training set. For ImageBind, the tactile encoder is aligned with the image encoder, and for LanguageBind, it is aligned with the language encoder, which is the same as the original training process. For training the connection between modalities, we employ multi-layer perceptrons to realize the variational information bottleneck framework. Moreover, to obtain the optimal classifier from each in-modality, we utilize the extracted features and train a linear layer as classification heads.

Setup. We consider two scenarios of OOM generalization: For the semi-supervised case, we sample 10% of the training data as labeled data with each class having a balanced number of labels. For the unsupervised case, we have no labels at all. For selecting the number of anchor points, we choose the same number of examples for the warm-up and training phases, which is 10% of the total training set. To train the OOM learner, we use the Adam optimizer with an initial learning rate of $1e-3$ with weight decay $1e-5$, and train for 50 epochs.

Baseline methods. Since there is no existing baseline method to compare with under our setting, we implement four methods for comparison, namely: Random where the model is randomly initialized, ERM where only labeled data is used to minimize the empirical risk, EntMin [351] which minimize the entropy of unlabeled data meanwhile conduct ERM, SSL which conducts self-supervised learning using Gaussian noise perturbation on the input, and MoCo [352] which updates model parameters with ensembling and meanwhile conducts

TABLE 6.2. Classification performance comparison of different methods across multiple datasets with different OOM modalities.

Setting	IM Perceptor	Method	TVL			LLVIP			NYU-D			VGGS		
			RGB	**Lan**	**Tac**	**RGB**	**Lan**	**The**	**RGB**	**Dep**	**Lan**	**Aud**	**Vid**	**Lan**
Semi-Supervised	ImageBind	Random	0.4	0.3	0.2	48.2	47.3	51.0	10.2	11.3	10.2	0.3	0.3	0.3
		ERM	23.1	19.5	22.7	54.6	53.1	54.1	45.2	44.5	38.1	9.3	10.2	8.4
		EntMin	24.0	21.8	23.6	56.7	57.0	55.4	48.0	46.3	39.3	10.5	13.3	8.9
		COX	**31.2**	**25.3**	**26.5**	**59.2**	**58.3**	**58.3**	**52.3**	**50.7**	**44.2**	**16.8**	**18.4**	**11.7**
		aligned	79.5	29.8	35.8	65.4	61.8	63.4	61.8	54.0	52.7	27.8	29.3	19.1
	LanguageBind	Random	0.4	0.3	0.2	48.2	47.3	51.0	10.2	11.3	10.2	0.3	0.3	0.3
		ERM	23.6	20.1	22.6	56.5	54.9	58.3	44.8	44.5	39.9	9.8	13.7	9.9
		EntMin	25.7	23.1	25.1	59.8	60.0	62.2	49.4	47.3	42.7	11.9	14.5	12.8
		COX	**33.5**	**26.3**	**27.3**	**61.2**	**62.3**	**66.4**	**58.8**	**53.5**	**48.4**	**18.3**	**22.1**	**13.4**
		aligned	81.6	31.2	38.3	74.1	73.2	87.2	68.6	65.1	57.7	38.6	32.5	20.9
Unsupervised	ImageBind	Random	0.4	0.3	0.2	48.2	47.3	51.0	10.2	11.3	10.2	0.3	0.3	0.3
		SSL	6.3	4.3	5.1	52.3	56.1	52.4	14.6	13.6	18.9	2.5	6.9	3.8
		COX	**18.9**	**15.4**	**17.1**	**54.8**	**57.2**	**53.8**	**21.7**	**22.0**	**19.5**	**9.3**	**10.2**	**10.5**
		aligned	79.5	29.8	35.8	65.4	61.8	63.4	61.8	54.0	52.7	27.8	29.3	19.1
	LanguageBind	Random	0.4	0.3	0.2	48.2	47.3	51.0	10.2	11.3	10.2	0.3	0.3	0.3
		SSL	6.8	6.5	5.1	54.6	57.8	53.8	16.9	18.1	16.3	7.2	5.6	4.8
		COX	**19.3**	**19.2**	**18.6**	**55.0**	**56.4**	**55.7**	**24.5**	**23.1**	**20.4**	**10.0**	**11.6**	**10.4**
		aligned	81.6	31.2	38.3	74.1	73.2	87.2	68.6	65.1	57.7	38.6	32.5	20.9

contrastive learning. Note that we use MoCo for comparison for the retrieval task in Table 6.3 because it is not for classification, and it is combined with EntMin in the semi-supervised case. Moreover, we use a pre-trained encoder as an upper-limit baseline "aligned". Next, we carefully compare the performance of our COX to these baseline methods.

6.4.2 Performance Comparison

For performance comparisons, we conduct classification and cross-modal retrieval to validate the proposed COX. There are seven modalities are considered, namely RGB image, language, tactile, thermal, depth, audio, and video which are simplified as RGB, Lan, Tac, The, Dep, Aud, and Vid, respectively. For each column, we choose one modality as OOM data, the rest modalities are selected IM data. For the retrieval task, we report the recall rate in both top 1 (R@1) and top 5 (R@5). The results are shown in Tables 6.2 and 6.3. We can see that the proposed COX clearly shows the best performance in both scenarios. Specifically, COX can achieve more than 5% performance improvement for most of the

TABLE 6.3. Cross-modal retrieval performance comparison of different methods across multiple datasets with different OOM modalities.

Setting	IM Perceptor	Method	MSR-VTT						MOSEI					
			Aud		Lan		Vid		Aud		Lan		Vid	
			R@1	R@5	R@1	R@5	R@1	R@5	R@1	R@5	R@1	R@5	R@1	R@5
Semi-Supervised	ImageBind	Random	5.4	25.1	5.0	25.4	5.4	24.2	14.3	42.5	14.4	42.8	14.1	42.1
		ERM	15.6	30.3	16.1	35.2	18.5	38.2	28.0	45.3	29.3	47.1	33.4	48.2
		EntMin	18.5	32.4	19.2	38.5	21.0	39.4	29.6	46.7	32.0	48.7	35.4	50.5
		MoCo	20.5	33.9	21.1	38.9	23.4	43.5	30.1	47.3	32.7	50.1	36.2	51.0
		COX	**23.3**	**35.8**	**23.4**	**39.1**	**26.5**	**48.8**	**32.4**	**48.0**	**33.8**	**50.4**	**38.8**	**53.7**
		Aligned	35.5	51.5	32.3	52.4	36.8	61.8	42.9	66.4	48.2	69.4	50.5	71.6
	LanguageBind	Random	5.2	24.3	5.4	25.1	5.0	25.6	13.5	43.1	14.2	42.7	14.6	41.9
		ERM	16.3	31.1	16.5	36.2	18.7	37.9	27.3	45.5	28.4	47.6	33.4	49.3
		EntMin	19.6	33.4	19.8	38.6	22.4	37.9	30.2	45.5	33.5	49.0	36.0	49.7
		MoCo	21.1	34.8	20.9	39.2	24.5	38.6	31.1	46.7	34.5	**50.5**	37.0	51.7
		COX	**25.2**	**36.0**	**24.1**	**40.0**	**28.7**	**49.5**	**34.6**	**49.8**	**34.6**	50.2	**39.2**	**55.4**
		Aligned	42.0	53.6	38.8	58.6	44.8	70.0	44.6	68.9	49.5	67.4	51.1	68.3
Unsupervised	ImageBind	Random	5.4	25.1	5.0	25.4	5.4	24.2	14.3	42.5	14.4	42.8	14.1	42.1
		SSL	8.9	28.4	9.3	28.1	10.1	29.5	17.4	48.8	16.2	45.2	16.0	45.0
		MoCo	9.2	28.9	9.5	28.4	10.6	30.0	17.8	50.3	16.6	45.8	17.1	44.4
		COX	**13.5**	**30.4**	**16.5**	**32.4**	**15.2**	**34.8**	**20.8**	**53.7**	**18.7**	**46.7**	**18.2**	**48.9**
		Aligned	35.5	51.5	32.3	52.4	36.8	61.8	42.9	66.4	48.2	69.4	50.5	71.6
	LanguageBind	Random	5.2	24.3	5.4	25.1	5.0	25.6	13.5	43.1	14.2	42.7	14.6	41.9
		SSL	9.2	28.9	11.0	28.8	10.3	28.7	18.0	48.9	18.4	45.0	17.8	45.6
		MoCo	9.6	29.4	11.1	28.5	11.0	29.3	18.8	50.7	18.5	45.2	18.0	45.5
		COX	**14.8**	**31.1**	**18.4**	**34.4**	**15.4**	**35.0**	**23.1**	**52.8**	**19.4**	**47.2**	**20.4**	**49.9**
		Aligned	42.0	53.6	38.8	58.6	44.8	70.0	44.6	68.9	49.5	67.4	51.1	68.3

OOM setting, which justifies that leveraging the knowledge from IM perceptors can indeed help OOM generalization compared to using OOM data alone. Moreover, even though the performance is relatively limited compared to the fully pre-trained baseline under the unsupervised case, considering it is an extremely challenging setting, we can still largely improve the performance for over 10% compared to the Random baseline, which demonstrates that the unsupervised OOM generalization is indeed learnable further leads to a novel research direction for improving the generalization performance. Additionally, note that the performance of COX is affected by the quality of IM perceptors, as using LanguageBind shows relatively higher performance compared to using ImageBind. Thus, it would be potentially helpful to leverage sophisticated IM perceptors to benefit the generalization performance.

6.4.3 Empirical Analysis

To provide an intuitive justification for the proposed method, here we conduct empirical analyses using the MSR-VTT dataset on various OOM scenarios and modalities.

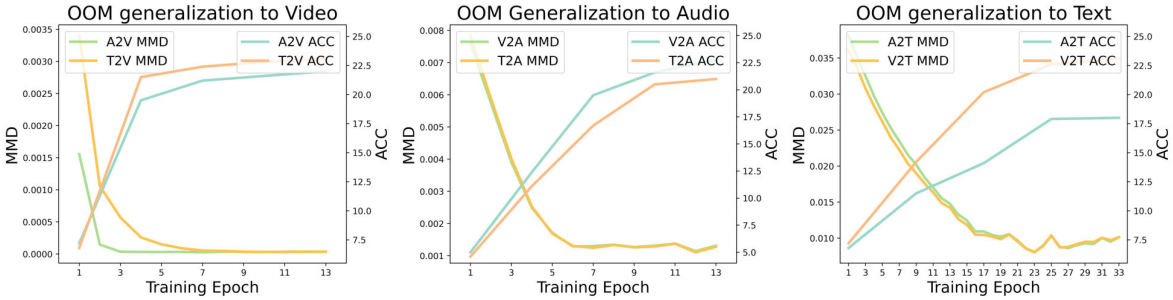


FIGURE 6.5. Connection effect on maximum mean discrepancy and accuracy across modalities.

Connection mitigates modality gap. To understand the performance of our VIB-based connection learning, here we show its effect on generalization out-of-modal. Specifically, during connection training, we compute the maximum mean discrepancy (MMD) between the mapping of each IM data and the OOM data. Meanwhile, we evenly select 6 points during the training and extract the IM mappings which are used to learn a classification head as the optimal classifier. Based on our theoretical result, we apply the classifiers to OOM data and compute their accuracies, as shown in Figure 6.5. We can see that as training goes on, the MMD between each IM mapping and OOM data is decreasing and the corresponding accuracy is increasing, which shows that: (1) our connection can indeed close the modality gap between their features and (2) as the mappings of IM data getting close to OOM data, the optimal classifier shows better classification results on OOM data, which benefits the knowledge transfer from known modalities to unknown ones.

Modality disagreement identifies uncertainty. To understand the effect of modality disagreement, we analyze the semi-supervised scenario by training the OOM learner to use only labeled data for 10 epochs. Then, we leverage the modality disagreement criteria to separate OOM data into those with disagreement and agreement and show their prediction accuracies

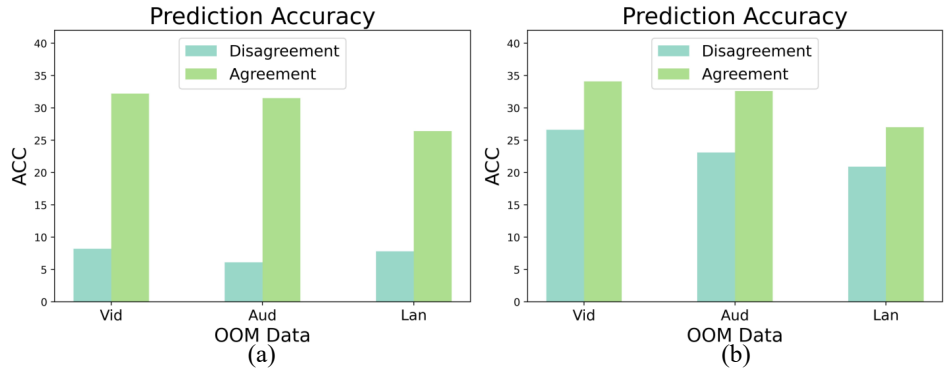


FIGURE 6.6. Prediction accuracy of OOM data with modality disagreement and modalities agreement, respectively. (a) Before exploration. (b) After exploration.

TABLE 6.4. Ablation study on various settings.

Setting	Ablation	MSR-VTT R@1		
		Aud	Lan	Vid
Semi	w/o connection	8.7	7.9	10.3
	w/o exploration	16.4	16.5	18.8
	COX	**25.2**	**24.1**	**40.0**
Unsup.	w/o warm-up	7.4	11.5	10.5
	COX	**14.8**	**18.4**	**15.4**

in Figure 6.6 (a). We can see that the accuracy for OOM data with disagreement is significantly lower than those with agreement, meaning that the prediction uncertainty, i.e., data with low accuracy, is effectively identified by the modality disagreement.

Modality agreement alleviates uncertainty. Further, we conduct training by following the procedure proposed in Section 6.3.3 and again show the accuracies of OOM data with disagreement and agreement in Figure 6.6 (b). We can see that the performance gap between the two types of data is largely mitigated, which justifies the methodology of exploring OOM data using the guidance of modality agreement. As a result, we can achieve almost comparable performance on both types of data, benefiting the overall generalization.

Ablation study. Additionally, we conduct an ablation study to justify the effect of our methodology. Specifically, we consider three ablations: (1) "w/o connection" where we remove the connection and directly apply the modality disagreement criteria on the original features

of IM data and OOM data, (2) "w/o exploration" where we only leverage the OOM data with agreement for training, (3) For unsupervised scenario, we consider "w/o warm-up" where we do not conduct the warm-up phase and directly training the model. The results in Table 6.4 show that all modules are essential for achieving effective OOM generalization. Specifically, the connection is vital for the knowledge transduction of IM data to OOM data, without which the generalization performance is largely degraded. The conclusion is consistent with the connection analysis where directly applying optimal classifiers across modalities leads to poor accuracy. Moreover, removing exploration also hinders the performance because the uniqueness of OOM data is largely ignored. Additionally, we find that the warm-up phase is essential for the unsupervised case. As initialized models have no classification capability, we need pre-training to form basic feature clusters that are consistent with IM data, further enabling effective OOM generalization.

Discussion on computational efficiency. Note that we conduct the feature connection mostly on the feature space, the computational cost of training VIB framework work is quite acceptable. The main cost is training the OOM learner which is the basic training with cross-entropy loss optimization and can be implemented on a single NVIDIA 3090/4090 GPU.

6.5 Conclusion and Limitation

In this Chapter, we study a novel and promising research direction dubbed Out-of-Modal (OOM) Generalization which aims to leverage knowledge from existing modalities to generalize to an unknown modality without instance-level correspondence. We consider two scenarios where there are a few correspondences and there is no correspondence, i.e., semi-supervised and unsupervised cases, respectively. To tackle these problems, we propose a Connect&Explore (COX) method which first learns connections across modalities to extract common knowledge and then explores the unique knowledge of OOM data based on modality disagreement. Extensive experiments are conducted to justify the proposed method and intuitive insights are provided to inspire future studies. However, our research is limited to several aspects which we hope to address in the future. First, although challenging as it

is, the performance is relatively limited compared to fully-aligned models, which requires more investigations to enhance generalization. Second, our OOM generalization is mostly conducted within the modalities from the same dataset. In the future, we hope to discover scenarios where the OOM data is from a different dataset with a large modality gap.

CHAPTER 7

Proofs and Theoretical Analyses

In this Chapter, we provide proofs and theoretical analyses, including the convergence analysis and theorem for sharpness-based worst-case optimization in Chapter 3, analysis and proposition for the pruning strategy in Chapter 4, theoretical proof for achieving optimal prediction via In-Context Learning in Chapter 5, and theoretical justification for our modality disagreement in Chapter 6.

7.1 Proof for SharpDRO

This section provides the proof of Theorem 1. We first give some notations before we start our proof for the convergence.

(1) We denote the loss expectation as $\mathbb{L}(\theta, \omega) := \mathbb{E}_{(x,y)\sim Q}\mathcal{L}(\theta, \omega; (x, y))$, and so as the SAM function that $\mathbb{R}(\theta, \omega) = \mathbb{E}_{(x,y)\sim Q}R(\theta, \omega; (x, y))$. So our objective can be turned into: $\min_\theta\{\max_\omega \mathbb{L}(\theta, \omega)\} + \mathbb{R}(\theta, \omega)$. And recalling our SharpDRO algorithm, we restate the meaning of the parameters: the model is parameterized by θ and ω means the weighted sampling.

(2) κ is the condition number that $\kappa = \frac{l}{\mu}$, where l is the Lipschitz-smoothness in Assumption 10 and μ means the PL condition in Assumption 12.

(3) We define $\mathbb{L}^*(\theta) = \max_\omega \mathbb{L}(\theta, \omega)$ and $\omega^*(\theta) = arg\max_\omega \mathbb{L}(\theta, \omega)$.

7.1.1 Update Rule

Before our theoretical analyses, we need to make the update rule for each variable explicit. We have to pay attention to the fact that our algorithm is stochastic that we can not directly get the real value of the gradient $\nabla\mathbb{L}(\theta,\omega)$, rather we estimate it by batches of samples $g_\theta(\theta,\omega) = \frac{1}{M}\sum_{i=1}^{M}\frac{\partial\mathcal{L}}{\partial\theta}(\theta,\omega;(x_i,y_i))$ and $g_\omega(\theta,\omega) = \frac{1}{M}\sum_{i=1}^{M}\frac{\partial\mathcal{L}}{\partial\omega}(\theta,\omega;(x_i,y_i))$, who hold some properties we will introduce in Assumption 8. So the optimization iteration is executed as follows in reality:

$$\begin{aligned}\theta_{t+1} &= \theta_t - \eta_\theta g_\theta(\theta_t + \rho g_\theta(\theta_t,\omega_t),\omega_t);\\ \omega_{t+1} &= \omega_t + \eta_\omega\nabla_\omega g_\omega(\theta_t,\omega_t).\end{aligned} \tag{7.1}$$

We further give a notation for brief that $\theta_{t+1/2} \triangleq \theta_t + \rho g_\theta(\theta_t,\omega_t)$, so the update for θ can be simplified as: $\theta_{t+1} = \theta_t - \eta_\theta g_\theta(\theta_{t+1/2},\omega_t)$.

7.1.2 Assumptions

We also have to make some necessary assumptions on our problem setting for this convergence proof:

ASSUMPTION 8 (Bounded variance). *The unbiased estimation about the gradient of the loss function also has bounded variance that:*

$$\mathbb{E}_{(x,y)\sim Q}[\frac{\partial\mathcal{L}}{\partial\theta}(\theta,\omega;(x,y))] = \nabla_\theta\mathbb{L}(\theta,\omega),\quad \mathbb{E}_{(x,y)\sim Q}\|\frac{\partial\mathcal{L}}{\partial\theta}(\theta,\omega;(x,y)) - \nabla_\theta\mathbb{L}(\theta,\omega)\|^2 \le \sigma^2;$$
$$\mathbb{E}_{(x,y)\sim Q}[\frac{\partial\mathcal{L}}{\partial\omega}(\theta,\omega;(x,y))] = \nabla_\omega\mathbb{L}(\theta,\omega),\quad \mathbb{E}_{(x,y)\sim Q}\|\frac{\partial\mathcal{L}}{\partial\omega}(\theta,\omega;(x,y)) - \nabla_\omega\mathbb{L}(\theta,\omega)\|^2 \le \sigma^2.$$

REMARK 9. *Since g_θ and g_ω are the averaged samples that: $g_\theta = \frac{1}{M}\sum_{i=1}^{M}\frac{\partial\mathcal{L}}{\partial\theta}(\theta,\omega;(x_i,y_i))$ and $g_\omega = \frac{1}{M}\sum_{i=1}^{M}\frac{\partial\mathcal{L}}{\partial\omega}(\theta,\omega;(x_i,y_i))$ respectively, they also have the unbiased property and have bounded variance:*

$$\mathbb{E}_{(x,y)\sim Q}[g_\theta(\theta,\omega;(x,y))] = \nabla_\theta\mathbb{L}(\theta,\omega),\quad \mathbb{E}_{(x,y)\sim Q}\|g_\theta(\theta,\omega;(x,y)) - \nabla_\theta\mathbb{L}(\theta,\omega)\|^2 \le \frac{\sigma^2}{M};$$
$$\mathbb{E}_{(x,y)\sim Q}[g_\omega(\theta,\omega;(x,y))] = \nabla_\omega\mathbb{L}(\theta,\omega),\quad \mathbb{E}_{(x,y)\sim Q}\|g_\omega(\theta,\omega;(x,y)) - \nabla_\omega\mathbb{L}(\theta,\omega)\|^2 \le \frac{\sigma^2}{M}.$$

ASSUMPTION 10 (Lipschitz smooth). *$\mathcal{L}(\theta, \omega; (x, y))$ is differential and l-Lipschitz smooth for every given sample (x, y):*

$$\|\nabla_\theta \mathcal{L}(\theta_1, \omega; (x, y)) - \nabla_\theta \mathcal{L}(\theta_2, \omega; (x, y))\| \leq l\|\theta_1 - \theta_2\|, \quad \forall \omega, (x, y);$$

$$\|\nabla_\omega \mathcal{L}(\theta, \omega_1; (x, y)) - \nabla_\omega \mathcal{L}(\theta, \omega_2; (x, y))\| \leq l\|\omega_1 - \omega_2\|, \quad \forall \theta, (x, y).$$

REMARK 11. *So the expectation function $\mathbb{L}$ also have the Lipschitz smooth property that:*

$$\|\nabla_\theta \mathbb{L}(\theta_1, \omega) - \nabla_\theta \mathbb{L}(\theta_2, \omega)\| \leq \mathbb{E}\|\nabla_\theta \mathcal{L}(\theta_1, \omega; (x, y)) - \nabla_\theta \mathcal{L}(\theta_2, \omega; (x, y))\| \leq l\|\theta_1 - \theta_2\|, \forall \omega;$$

$$\|\nabla_\omega \mathbb{L}(\theta, \omega_1) - \nabla_\omega \mathbb{L}(\theta, \omega_2)\| \leq \mathbb{E}\|\nabla_\omega \mathcal{L}(\theta, \omega_1; (x, y)) - \nabla_\omega \mathcal{L}(\theta, \omega_2; (x, y))\| \leq l\|\omega_1 - \omega_2\|, \forall \theta.$$

ASSUMPTION 12 (PL condition). *The loss function $\mathbb{L}(\theta, \cdot)$ satisfies PL condition on every given θ, i.e., there exists $\mu > 0$ such that $\|\nabla_\omega \mathbb{L}(\theta, \omega)\|^2 \geq 2\mu[\max_\omega \mathbb{L}(\theta, \omega) - \mathbb{L}(\theta, \omega)], \forall \theta, \omega$.*

7.1.3 Useful Lemmas

In this part, we will prove some necessary lemmas for us to prove the convergence bound. And we will give the definition of the stationary point of our problem at the beginning.

DEFINITION 13 (Stationary measure). *θ is defined as the ϵ-stationary point of our problem if $\mathbb{E}\|\nabla \mathbb{L}^*(\theta)\| \leq \epsilon$ for any $\epsilon \geq 0$.*

REMARK 14. *For minmax problem, there are usually two ways to measure the stationary point. One is measured two-side when $\mathbb{E}\|\nabla_\theta \mathbb{L}(\theta, \omega)\| \leq \epsilon$ and $\mathbb{E}\|\nabla_\omega \mathbb{L}(\theta, \omega)\| \leq \epsilon$, we claim (θ, ω) is the (ϵ, ϵ)-stationary point. It has been proved in [353] that these two measures can be translated into each other when $\mathbb{L}^*$ is smooth which will be shown in Lemma 2. But what we compute is the model parameter θ using the algorithm SharpDRO. So we choose the measure by $\mathbb{E}\|\mathbb{L}^*(\theta)\|$ here.*

LEMMA 2. *[354] Under Assumption 10 and 12, $\mathbb{L}^*(\theta)$ is $(l + \frac{l^2}{2\mu})$-Lipschitz smooth with the gradient:*

$$\nabla_\theta \mathbb{L}^*(\theta, \omega) = \nabla_\theta \mathbb{L}(\theta, \omega^*(\theta)).$$

LEMMA 3. *[354] Under Assumption 10 and 12, $\omega^*(\cdot)$ is smooth about its variable:*

$$\|\omega^*(\theta_1) - \omega^*(\theta_2)\| \leq \frac{l}{2\mu}\|\theta_1 - \theta_2\|, \quad \forall \theta_1, \theta_2.$$

LEMMA 4. *We give an estimation that:*

$$\mathbb{E}\|g_\theta(\theta_{t+1/2},\omega_t)\|^2 \leq (4\rho^2 l^2 + 2\rho l + 2)\mathbb{E}\|\nabla_\theta \mathbb{L}(\theta_t,\omega_t)\|^2 + (5\rho^2 l^2 + 2)\frac{\sigma^2}{M}. \tag{7.2}$$

PROOF.

$$\mathbb{E}\|g_\theta(\theta_{t+1/2},\omega_t)\|^2 = -\mathbb{E}\|\nabla_\theta \mathbb{L}(\theta_t,\omega_t)\|^2 + \mathbb{E}\|g_\theta(\theta_{t+1/2},\omega_t) - \nabla_\theta \mathbb{L}(\theta_t,\omega_t)\|^2 \tag{7.3}$$

$$+ 2\mathbb{E}\langle g_\theta(\theta_{t+1/2},\omega_t), \nabla_\theta \mathbb{L}(\theta_t,\omega_t)\rangle. \tag{7.4}$$

For the cross-product term, we divide it as follows:

$$\begin{aligned}
&\mathbb{E}\langle g_\theta(\theta_{t+1/2},\omega_t), \nabla_\theta \mathbb{L}(\theta_t,\omega_t)\rangle \\
=\ &\mathbb{E}\langle g_\theta(\theta_{t+1/2},\omega_t) - g_\theta(\theta_t + \rho\nabla_\theta \mathbb{L}(\theta_t,\omega_t),\omega_t), \nabla_\theta \mathbb{L}(\theta_t,\omega_t)\rangle \\
&+ \mathbb{E}\langle g_\theta(\theta_t + \rho\nabla_\theta \mathbb{L}(\theta_t,\omega_t),\omega_t), \nabla_\theta \mathbb{L}(\theta_t,\omega_t)\rangle \\
=\ &\mathbb{E}\langle \nabla_\theta \mathbb{L}(\theta_{t+1/2},\omega_t) - \nabla_\theta \mathbb{L}(\theta_t + \rho\nabla_\theta \mathbb{L}(\theta_t,\omega_t),\omega_t), \nabla_\theta \mathbb{L}(\theta_t,\omega_t)\rangle \\
&+ \mathbb{E}\langle \nabla_\theta \mathbb{L}(\theta_t + \rho\nabla_\theta \mathbb{L}(\theta_t,\omega_t),\omega_t), \nabla_\theta \mathbb{L}(\theta_t,\omega_t)\rangle \overset{(i)}{\leq} \frac{1}{2}\mathbb{E}\|\nabla_\theta \mathbb{L}(\theta_{t+1/2},\omega_t) \\
&- \nabla_\theta \mathbb{L}(\theta_t + \rho\nabla_\theta \mathbb{L}(\theta_t,\omega_t),\omega_t)\|^2 + \frac{1}{2}\mathbb{E}\|\nabla_\theta \mathbb{L}(\theta_t,\omega_t)\|^2 + \mathbb{E}\|\nabla_\theta \mathbb{L}(\theta_t,\omega_t)\|^2 \\
&+ \mathbb{E}\langle \nabla_\theta \mathbb{L}(\theta_t + \rho\nabla_\theta \mathbb{L}(\theta_t,\omega_t),\omega_t) - \nabla_\theta \mathbb{L}(\theta_t,\omega_t), \nabla_\theta \mathbb{L}(\theta_t,\omega_t)\rangle \\
\overset{(ii)}{\leq}\ &\frac{\rho^2 l^2}{2}\mathbb{E}\|g_\theta(\theta_t,\omega_t) - \nabla_\theta \mathbb{L}(\theta_t,\omega_t)\|^2 + \frac{3}{2}\mathbb{E}\|\nabla_\theta \mathbb{L}(\theta_t,\omega_t)\|^2 + \rho l\mathbb{E}\|\nabla_\theta \mathbb{L}(\theta_t,\omega_t)\|^2 \\
\overset{(iii)}{\leq}\ &(\rho l + \frac{3}{2})\mathbb{E}\|\nabla_\theta \mathbb{L}(\theta_t,\omega_t)\|^2 + \frac{\rho^2 l^2 \sigma^2}{2M},
\end{aligned} \tag{7.5}$$

where the inequality (i) is due to the Cauchy-Schwarz inequality; the inequality (ii) is because of the Lipschitz-smoothness of $\mathbb{L}$ that:

$$\mathbb{E}\|\nabla_\theta \mathbb{L}(\theta_{t+1/2},\omega_t) - \nabla_\theta \mathbb{L}(\theta_t + \rho\nabla_\theta \mathbb{L}(\theta_t,\omega_t),\omega_t)\|^2 \leq l^2\mathbb{E}\|\theta_{t+1/2} - \theta_t - \rho\nabla_\theta \mathbb{L}(\theta_t,\omega_t)\|^2, \tag{7.6}$$

and the property of Lipschitz-smoothness that:

$$\langle \nabla_\theta \mathbb{L}(\theta_t + \rho \nabla_\theta \mathbb{L}(\theta_t, \omega_t), \omega_t) - \nabla_\theta \mathbb{L}(\theta_t, \omega_t), \nabla_\theta \mathbb{L}(\theta_t, \omega_t) \rangle = \tag{7.7}$$

$$\frac{1}{\rho} \langle \nabla_\theta \mathbb{L}(\theta_t + \rho \nabla_\theta \mathbb{L}(\theta_t, \omega_t), \omega_t) - \nabla_\theta \mathbb{L}(\theta_t, \omega_t), \rho \nabla_\theta \mathbb{L}(\theta_t, \omega_t) \rangle \leq \frac{l}{\rho} \|\rho \nabla_\theta \mathbb{L}(\theta_t, \omega_t)\|^2, \tag{7.8}$$

and the inequality (iii) makes use of the Assumption 8.

As for the second term, we have:

$$\begin{aligned}
&\mathbb{E}\|g_\theta(\theta_{t+1/2}, \omega_t) - \nabla_\theta \mathbb{L}(\theta_t, \omega_t)\|^2 \\
&\leq 2\mathbb{E}\|g_\theta(\theta_{t+1/2}, \omega_t) - \nabla_\theta \mathbb{L}(\theta_{t+1/2}, \omega_t)\|^2 + 2\mathbb{E}\|\nabla_\theta \mathbb{L}(\theta_{t+1/2}, \omega_t) - \nabla_\theta \mathbb{L}(\theta_t, \omega_t)\|^2 \\
&\leq \frac{2\sigma^2}{M} + 2l^2 \mathbb{E}\|\theta_{t+1/2} - \theta_t\|^2 \\
&= \frac{2\sigma^2}{M} + 2\rho^2 l^2 \mathbb{E}\|g_\theta(\theta_t, \omega_t)\|^2 \\
&\leq 2\frac{\sigma^2}{M}(2\rho^2 l^2 + 1) + 4\rho^2 l^2 \mathbb{E}\|\nabla_\theta \mathbb{L}(\theta_t, \omega_t)\|^2,
\end{aligned} \tag{7.9}$$

where the last inequality comes from the fact that:

$$\mathbb{E}\|g_\theta(\theta_t, \omega_t)\|^2 \leq 2\mathbb{E}\|g_\theta(\theta_t, \omega_t) - \nabla_\theta \mathbb{L}(\theta_t, \omega_t)\|^2 + 2\mathbb{E}\|\nabla_\theta \mathbb{L}(\theta_t, \omega_t)\|^2. \tag{7.10}$$

By combining the above inequalities, we can get:

$$\mathbb{E}\|g_\theta(\theta_{t+1/2}, \omega_t)\|^2 \leq (4\rho^2 l^2 + 2\rho l + 2)\mathbb{E}\|\nabla_\theta \mathbb{L}(\theta_t, \omega_t)\|^2 + (5\rho^2 l^2 + 2)\frac{\sigma^2}{M}. \tag{7.11}$$

□

LEMMA 5. *For the descending relationship of the function $\mathbb{L}^*$, we have:*

$$\begin{aligned}
\mathbb{E}[\mathbb{L}^*(\theta_{t+1})] \leq \mathbb{E}[\mathbb{L}^*(\theta_t)] &- \frac{\eta_\theta}{2}(1 - 5\rho l - 2L\eta_\theta(4\rho^2 l^2 + 2\rho l + 2))\mathbb{E}\|\nabla \mathbb{L}^*(\theta_t)\|^2 \\
&+ [\frac{\eta_\theta}{2}(1 + \frac{1}{2}\rho l) + L\eta_\theta^2(4\rho^2 l^2 + 2\rho l + 2)]\mathbb{E}\|\nabla \mathbb{L}^*(\theta_t) - \nabla_\theta \mathbb{L}(\theta_t, \omega_t)\|^2 \\
&+ (5\rho^2 l^2 + 2)\frac{L\eta_\theta^2 \sigma^2}{2M},
\end{aligned}$$

where we use the brief notation that $L = l + \frac{l\kappa}{2}$.

PROOF. Since $\mathbb{L}^*(\theta)$ is $(l + \frac{l\kappa}{2})$-Lipschitz smooth according to Lemma 2, we have:

$$\begin{aligned}\mathbb{L}^*(\theta_{t+1}) &\leq \mathbb{L}^*(\theta_t) + \langle \nabla \mathbb{L}^*(\theta_t), \theta_{t+1} - \theta_t \rangle + \frac{1}{2}(l + \frac{l\kappa}{2})\|\theta_{t+1} - \theta_t\|^2 \\ &= \mathbb{L}^*(\theta_t) - \eta_\theta \langle \nabla \mathbb{L}^*(\theta_t), g_\theta(\theta_{t+1/2}, \omega_t) \rangle + \frac{1}{2}(l + \frac{l\kappa}{2})\eta_\theta^2 \|g_\theta(\theta_{t+1/2}, \omega_t)\|^2.\end{aligned} \tag{7.12}$$

Taking expectation conditioned on (θ_t, ω_t) and we get:

$$\mathbb{E}[\mathbb{L}^*(\theta_{t+1})|\theta_t, \omega_t] \leq \tag{7.13}$$

$$\mathbb{L}^*(\theta_t) - \eta_\theta \langle \nabla \mathbb{L}^*(\theta_t), \nabla_\theta \mathbb{L}(\theta_{t+1/2}, \omega_t) \rangle + \frac{1}{2}(l + \frac{l\kappa}{2})\eta_\theta^2 \mathbb{E}[\|g_\theta(\theta_{t+1/2}, \omega_t)\|^2 | \theta_t, \omega_t]. \tag{7.14}$$

We again take expectation on both side on above inequality so we have:

$$\mathbb{E}[\mathbb{L}^*(\theta_{t+1})] \leq \mathbb{E}[\mathbb{L}^*(\theta_t)] - \eta_\theta \mathbb{E}\langle \nabla \mathbb{L}^*(\theta_t), \nabla_\theta \mathbb{L}(\theta_{t+1/2}, \omega_t) \rangle + \frac{1}{2}(l + \frac{l\kappa}{2})\eta_\theta^2 \mathbb{E}\|g_\theta(\theta_{t+1/2}, \omega_t)\|^2. \tag{7.15}$$

For the second term, we decompose it as follows:

$$\begin{aligned}&\mathbb{E}\langle \nabla \mathbb{L}^*(\theta_t), \nabla_\theta \mathbb{L}(\theta_{t+1/2}, \omega_t) \rangle \\ =\ &\mathbb{E}\langle \nabla \mathbb{L}^*(\theta_t), \nabla_\theta \mathbb{L}(\theta_t, \omega_t) + \nabla_\theta \mathbb{L}(\theta_{t+1/2}, \omega_t) - \nabla_\theta \mathbb{L}(\theta_t, \omega_t) \rangle \\ \geq\ &\mathbb{E}\langle \nabla \mathbb{L}^*(\theta_t), \nabla_\theta \mathbb{L}(\theta_t, \omega_t) \rangle - \mathbb{E}\|\nabla \mathbb{L}^*(\theta_t)\| \|\nabla_\theta \mathbb{L}(\theta_{t+1/2}, \omega_t) - \nabla_\theta \mathbb{L}(\theta_t, \omega_t)\| \\ \geq\ &\mathbb{E}\langle \nabla \mathbb{L}^*(\theta_t), \nabla_\theta \mathbb{L}(\theta_t, \omega_t) \rangle - \rho l \mathbb{E}\|\nabla \mathbb{L}^*(\theta_t)\| \|g_\theta(\theta_t, \omega_y)\| \\ \geq\ &\mathbb{E}\langle \nabla \mathbb{L}^*(\theta_t), \nabla \mathbb{L}^*(\theta_t) + \nabla_\theta \mathbb{L}(\theta_t, \omega_t) - \nabla \mathbb{L}^*(\theta_t) \rangle \\ &- \rho l \mathbb{E}\|\nabla \mathbb{L}^*(\theta_t)\| (\|\nabla_\theta \mathbb{L}(\theta_t, \omega_t)\| + \|g_\theta(\theta_t, \omega_t) - \nabla_\theta \mathbb{L}(\theta_t, \omega_t)\|) \\ \geq\ &\mathbb{E}\|\nabla \mathbb{L}^*(\theta_t)\|^2 - \frac{1}{2}\mathbb{E}\|\nabla \mathbb{L}^*(\theta_t)\|^2 - \frac{1}{2}\mathbb{E}\|\nabla_\theta \mathbb{L}(\theta_t, \omega_t) - \nabla \mathbb{L}^*(\theta_t)\|^2 \\ &- \rho l \mathbb{E}\|\nabla \mathbb{L}^*(\theta_t)\| \|\nabla_\theta \mathbb{L}(\theta_t, \omega_t)\| \\ &- \frac{1}{2}\rho l \mathbb{E}\|\nabla \mathbb{L}^*(\theta_t)\|^2 - \frac{1}{2}\rho l \mathbb{E}\|g_\theta(\theta_t, \omega_t) - \nabla_\theta \mathbb{L}(\theta_t, \omega_t)\|^2 \\ \geq\ &\frac{1 - \rho l}{2}\mathbb{E}\|\nabla \mathbb{L}^*(\theta_t)\|^2 - \frac{1}{2}\mathbb{E}\|\nabla_\theta \mathbb{L}(\theta_t, \omega_t) - \nabla \mathbb{L}^*(\theta_t)\|^2 \\ &- \rho l \mathbb{E}\|\nabla \mathbb{L}^*(\theta_t)\| \|\nabla_\theta \mathbb{L}(\theta_t, \omega_t)\| - \frac{\rho l \sigma^2}{2M}.\end{aligned} \tag{7.16}$$

We continue estimating the last term in above inequality 7.16

$$\begin{aligned}
&\mathbb{E}\|\nabla\mathbb{L}^*(\theta_t)\|\|\nabla_\theta\mathbb{L}(\theta_t,\omega_t)\| \\
&= \mathbb{E}\|\nabla\mathbb{L}^*(\theta_t)\|\|\nabla_\theta\mathbb{L}(\theta_t,\omega_t) - \nabla\mathbb{L}^*(\theta_t) + \nabla\mathbb{L}^*(\theta_t)\| \\
&\leq \mathbb{E}\|\nabla\mathbb{L}^*(\theta_t)\|^2 + \mathbb{E}\|\nabla\mathbb{L}^*(\theta_t)\|\|\nabla_\theta\mathbb{L}(\theta_t,\omega_t) - \nabla\mathbb{L}^*(\theta_t)\| \\
&\overset{(i)}{\leq} \mathbb{E}\|\nabla\mathbb{L}^*(\theta_t)\|^2 + \mathbb{E}\|\nabla\mathbb{L}^*(\theta_t)\|^2 + \frac{1}{4}\mathbb{E}\|\nabla_\theta\mathbb{L}(\theta_t,\omega_t) - \nabla\mathbb{L}^*(\theta_t)\|^2,
\end{aligned} \tag{7.17}$$

where the last inequality (i) is due to Young's inequality.

By combining inequality 7.15 with 7.17, we can get:

$$\begin{aligned}
&\mathbb{E}\langle\nabla\mathbb{L}^*(\theta_t), \nabla_\theta\mathbb{L}(\theta_{t+1/2},\omega_t)\rangle \\
&\geq \frac{1-\rho l}{2}\mathbb{E}\|\nabla\mathbb{L}^*(\theta_t)\|^2 - \frac{1}{2}\mathbb{E}\|\nabla_\theta\mathbb{L}(\theta_t,\omega_t) - \nabla\mathbb{L}^*(\theta_t)\|^2 \\
&\quad - 2\rho l\mathbb{E}\|\nabla\mathbb{L}^*(\theta_t)\|^2 - \frac{\rho l}{4}\mathbb{E}\|\nabla_\theta\mathbb{L}(\theta_t,\omega_t) - \nabla\mathbb{L}^*(\theta_t)\|^2 - \frac{\rho l\sigma^2}{2M} \\
&= \frac{1}{2}(1-5\rho l)\mathbb{E}\|\nabla\mathbb{L}^*(\theta_t)\|^2 - \frac{1}{2}(1+\frac{1}{2}\rho l)\mathbb{E}\|\nabla\mathbb{L}^*(\theta_t) - \nabla_\theta\mathbb{L}(\theta_t,\omega_t)\|^2 - \frac{\rho l\sigma^2}{2M}.
\end{aligned} \tag{7.18}$$

Finally, we combine inequality 7.15 with Lemma 4 and inequality 7.18:

$$\begin{aligned}
&\mathbb{E}[\mathbb{L}^*(\theta_{t+1})] \\
&\leq \mathbb{E}[\mathbb{L}^*(\theta_t)] - \frac{\eta_\theta}{2}(1-5\rho l)\mathbb{E}\|\nabla\mathbb{L}^*(\theta_t)\|^2 + \frac{\eta_\theta}{2}(1+\frac{1}{2}\rho l)\mathbb{E}\|\nabla\mathbb{L}^*(\theta_t) - \nabla_\theta\mathbb{L}(\theta_t,\omega_t)\|^2 \\
&\quad + \frac{1}{2}(l+\frac{l\kappa}{2})\eta_\theta^2((4\rho^2l^2+2\rho l+2)\mathbb{E}\|\nabla_\theta\mathbb{L}(\theta_t,\omega_t)\|^2 + (5\rho^2l^2+2)\frac{\sigma^2}{M}) \\
&\overset{(i)}{\leq} \mathbb{E}[\mathbb{L}^*(\theta_t)] - \frac{\eta_\theta}{2}(1-5\rho l - \eta_\theta(2l+l\kappa)(4\rho^2l^2+2\rho l+2))\mathbb{E}\|\nabla\mathbb{L}^*(\theta_t)\|^2 \\
&\quad + [\frac{\eta_\theta}{2}(1+\frac{1}{2}\rho l) + \eta_\theta^2(l+\frac{l\kappa}{2})(4\rho^2l^2+2\rho l+2)]\mathbb{E}\|\nabla\mathbb{L}^*(\theta_t) - \nabla_\theta\mathbb{L}(\theta_t,\omega_t)\|^2 \\
&\quad + \frac{1}{2}(l+\frac{l\kappa}{2})(5\rho^2l^2+2)\frac{\eta_\theta^2\sigma^2}{M},
\end{aligned} \tag{7.19}$$

where the last inequality (i) uses the Cauchy-Schwarz inequality that $\|\nabla_\theta\mathbb{L}(\theta_t,\omega_t)\|^2 \leq 2\|\nabla\mathbb{L}^*(\theta_t)\|^2 + 2\|\nabla_\theta\mathbb{L}(\theta_t,\omega_t) - \nabla\mathbb{L}^*(\theta_t)\|^2$. □

7.1.4 Theorem

THEOREM 15. *Under Assumption 8, 10, 12, and the learning rate satisfy that:*

$$\eta_\theta \leq \min\{\frac{1}{128\kappa^2 l}, \sqrt{\frac{M(\mathbb{E}[\mathbb{L}^*(\theta_0)] - \min_\theta \mathbb{E}[\mathbb{L}^*(\theta)])}{132T\kappa^4 l\sigma^2}}\}, \tag{7.20}$$

$\eta_\omega \leq 64\kappa^2\eta_\theta$ *and* $\rho \leq \frac{\eta_\theta}{2l}$*, we have the convergence bound for our problem:*

$$\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}\|\nabla\mathbb{L}^*(\theta_t)\|^2 \leq 320\sqrt{\frac{3\kappa^4 l(\mathbb{E}[\mathbb{L}^*(\theta_0)] - \min_\theta \mathbb{E}[\mathbb{L}^*(\theta)])\sigma^2}{11MT}} = \mathcal{O}(\frac{\kappa^2}{\sqrt{MT}}). \tag{7.21}$$

PROOF. First recall the descending relationship of the function $\mathbb{L}^*$ in Lemma 5:

$$\begin{aligned}\mathbb{E}[\mathbb{L}^*(\theta_{t+1})] \leq\ & \mathbb{E}[\mathbb{L}^*(\theta_t)] - \frac{\eta_\theta}{2}(1 - 5\rho l - 2L\eta_\theta(4\rho^2 l^2 + 2\rho l + 2))\mathbb{E}\|\nabla\mathbb{L}^*(\theta_t)\|^2 \\ & + [\frac{\eta_\theta}{2}(1 + \frac{1}{2}\rho l) + L\eta_\theta^2(4\rho^2 l^2 + 2\rho l + 2)]\mathbb{E}\|\nabla\mathbb{L}^*(\theta_t) - \nabla_\theta\mathbb{L}(\theta_t, \omega_t)\|^2 \\ & + (5\rho^2 l^2 + 2)\frac{L\eta_\theta^2\sigma^2}{2M}.\end{aligned} \tag{7.22}$$

Then, using the smoothness of the variables θ and ω respectively, we can get:

$$\mathbb{L}(\theta_{t+1}, \omega_t) \geq \mathbb{L}(\theta_t, \omega_t) + \langle\nabla_\theta\mathbb{L}(\theta_t, \omega_t), \theta_{t+1} - \theta_t\rangle - \frac{l}{2}\|\theta_{t+1} - \theta_t\|^2;$$

$$\mathbb{L}(\theta_{t+1}, \omega_{t+1}) \geq \mathbb{L}(\theta_{t+1}, \omega_t) + \langle\nabla_\omega\mathbb{L}(\theta_{t+1}, \omega_t), \omega_{t+1} - \omega_t\rangle - \frac{l}{2}\|\omega_{t+1} - \omega_t\|^2.$$

Taking expectation, we can get:

$$\begin{aligned}
\mathbb{E}[\mathbb{L}(\theta_{t+1},\omega_t)] &\geq \mathbb{E}[\mathbb{L}(\theta_t,\omega_t)] - \eta_\theta \mathbb{E}\langle \nabla_\theta \mathbb{L}(\theta_t,\omega_t), \nabla_\theta \mathbb{L}(\theta_{t+1/2},\omega_t)\rangle \\
&\quad - \frac{l\eta_\theta^2}{2}\mathbb{E}\|g_\theta(\theta_{t+1/2},\omega_t)\|^2 \\
&\geq \mathbb{E}[\mathbb{L}(\theta_t,\omega_t)] - \eta_\theta \mathbb{E}\|\nabla_\theta \mathbb{L}(\theta_t,\omega_t)\|^2 - \frac{\eta_\theta}{2}\mathbb{E}\|\nabla_\theta \mathbb{L}(\theta_t,\omega_t)\|^2 \\
&\quad - \frac{\eta_\theta}{2}\mathbb{E}\|\nabla_\theta \mathbb{L}(\theta_{t+1/2},\omega_t) - \nabla_\theta \mathbb{L}(\theta_t,\omega_t)\|^2 - \frac{l\eta_\theta^2}{2}\mathbb{E}\|g_\theta(\theta_{t+1/2},\omega_t)\|^2 \\
&\geq \mathbb{E}[\mathbb{L}(\theta_t,\omega_t)] - \frac{3\eta_\theta}{2}\mathbb{E}\|\nabla_\theta \mathbb{L}(\theta_t,\omega_t)\|^2 - \frac{l^2\rho^2\eta_\theta}{2}\mathbb{E}\|g_\theta(\theta_t,\omega_t)\|^2 \\
&\quad - \frac{l\eta_\theta^2}{2}\mathbb{E}\|g_\theta(\theta_{t+1/2},\omega_t)\|^2 \\
&\geq \mathbb{E}[\mathbb{L}(\theta_t,\omega_t)] - (\frac{3\eta_\theta}{2} + \frac{l^2\rho^2\eta_\theta}{2} + l\eta_\theta^2(2\rho^2l^2+\rho l+1))\mathbb{E}\|\nabla_\theta \mathbb{L}(\theta_t,\omega_t)\|^2 \\
&\quad - (\frac{l^2\rho^2\eta_\theta}{2} + \frac{l\eta_\theta^2}{2}(5\rho^2l^2+2))\frac{\sigma^2}{M};
\end{aligned}$$

$$\begin{aligned}
\mathbb{E}[\mathbb{L}(\theta_{t+1},\omega_{t+1})] &\geq \mathbb{E}[\mathbb{L}(\theta_{t+1},\omega_t)] + \eta_\omega \mathbb{E}\langle \nabla_\omega \mathbb{L}(\theta_{t+1},\omega_t), \nabla_\omega \mathbb{L}(\theta_t,\omega_t)\rangle - \frac{l\eta_\omega^2}{2}\mathbb{E}\|g_\omega(\theta_t,\omega_t)\|^2 \\
&\geq \mathbb{E}[\mathbb{L}(\theta_{t+1},\omega_t)] + \frac{\eta_\omega}{2}\mathbb{E}\|\nabla_\omega \mathbb{L}(\theta_t,\omega_t)\|^2 - \frac{\eta_\omega}{2}\mathbb{E}\|\nabla_\omega \mathbb{L}(\theta_{t+1},\omega_t) \\
&\quad - \nabla_\omega \mathbb{L}(\theta_t,\omega_t)\|^2 - \frac{l\eta_\omega^2}{2}\mathbb{E}\|g_\omega(\theta_t,\omega_t)\|^2 \\
&\geq \mathbb{E}[\mathbb{L}(\theta_{t+1},\omega_t)] + \frac{\eta_\omega}{2}\mathbb{E}\|\nabla_\omega \mathbb{L}(\theta_t,\omega_t)\|^2 - \frac{l\eta_\theta^2\eta_\omega}{2}\mathbb{E}\|g_\theta(\theta_{t+1/2},\omega_t)\|^2 \\
&\quad - \frac{l\eta_\omega^2}{2}\mathbb{E}\|g_\omega(\theta_t,\omega_t)\|^2 \\
&\geq \mathbb{E}[\mathbb{L}(\theta_{t+1},\omega_t)] + (\frac{\eta_\omega}{2} - \frac{l\eta_\omega^2}{2})\mathbb{E}\|\nabla_\omega \mathbb{L}(\theta_t,\omega_t)\|^2 \\
&\quad - (l\eta_\theta^2\eta_\omega(2\rho^2l^2+\rho l+1))\mathbb{E}\|\nabla_\theta \mathbb{L}(\theta_t,\omega_t)\|^2 - (\frac{l\eta_\omega^2}{2} + \frac{l\eta_\theta^2\eta_\omega}{2}(5\rho^2l^2+2))\frac{\sigma^2}{M}.
\end{aligned} \tag{7.23}$$

Then we construct a potential function in the same way as [353]:

$$V_t = V(\theta_t,\omega_t) = \mathbb{L}^*(\theta_t) + \alpha[\mathbb{L}^*(\theta_t) - \mathbb{L}(\theta_t,\omega_t)],$$

where $\alpha > 0$ is a preset parameter. Then we come to evaluate the descending relationship of the potential function V_t.

Combining the above inequalities, we get the potential function descending relationship:

$$
\begin{aligned}
&\mathbb{E}[V_{t+1}] - \mathbb{E}[V_t] \\
&= (1+\alpha)(\mathbb{E}[\mathbb{L}^*(\theta_{t+1})] - \mathbb{E}[\mathbb{L}^*(\theta_t)]) - \alpha(\mathbb{E}[\mathbb{L}(\theta_{t+1},\omega_{t+1})] - \mathbb{E}[\mathbb{L}(\theta_t,\omega_t)]) \\
&\leq (1+\alpha)\{-\frac{\eta_\theta}{2}(1-5\rho l - 2L\eta_\theta(4\rho^2 l^2 + 2\rho l + 2))\mathbb{E}\|\nabla\mathbb{L}^*(\theta_t)\|^2 \\
&\quad + [\frac{\eta_\theta}{2}(1+\frac{1}{2}\rho l) + L\eta_\theta^2(4\rho^2 l^2 + 2\rho l + 2)]\mathbb{E}\|\nabla\mathbb{L}^*(\theta_t) - \nabla_\theta\mathbb{L}(\theta_t,\omega_t)\|^2 \qquad (7.24) \\
&\quad + (5\rho^2 l^2 + 2)\frac{L\eta_\theta^2\sigma^2}{2M}\} - \alpha\{-(\frac{3\eta_\theta}{2} + \frac{l^2\rho^2\eta_\theta}{2} + l\eta_\theta^2(2\rho^2 l^2 + \rho l + 1))\mathbb{E}\|\nabla_\theta\mathbb{L}(\theta_t,\omega_t)\|^2 \\
&\quad - (\frac{l^2\rho^2\eta_\theta}{2} + \frac{l\eta_\theta^2}{2}(5\rho^2 l^2 + 2))\frac{\sigma^2}{M} + (\frac{\eta_\omega}{2} - \frac{l\eta_\omega^2}{2})\mathbb{E}\|\nabla_\omega\mathbb{L}(\theta_t,\omega_t)\|^2 \\
&\quad - (l\eta_\theta^2\eta_\omega(2\rho^2 l^2 + \rho l + 1))\mathbb{E}\|\nabla_\theta\mathbb{L}(\theta_t,\omega_t)\|^2 - (\frac{l\eta_\omega^2}{2} + \frac{l\eta_\theta^2\eta_\omega}{2}(5\rho^2 l^2 + 2))\frac{\sigma^2}{M}\} \qquad (7.25) \\
&= -\frac{\eta_\theta}{2}(1+\alpha)(1-5\rho l - 2L\eta_\theta(4\rho^2 l^2 + 2\rho l + 2))\mathbb{E}\|\nabla\mathbb{L}^*(\theta_t)\|^2 \\
&\quad + (1+\alpha)(\frac{\eta_\theta}{2}(1+\frac{1}{2}\rho l) + L\eta_\theta^2(4\rho^2 l^2 + 2\rho l + 2))\mathbb{E}\|\nabla\mathbb{L}^*(\theta_t) - \nabla_\theta\mathbb{L}(\theta_t,\omega_t)\|^2 \\
&\quad + \alpha[(\frac{3\eta_\theta}{2} + \frac{l^2\rho^2\eta_\theta}{2} + l\eta_\theta^2(2\rho^2 l^2 + \rho l + 1)) + l\eta_\theta^2\eta_\omega(2\rho^2 l^2 + \rho l + 1)]\mathbb{E}\|\nabla_\theta\mathbb{L}(\theta_t,\omega_t)\|^2 \\
&\quad - \alpha(\frac{\eta_\omega}{2} - \frac{l\eta_\omega^2}{2})\mathbb{E}\|\nabla_\omega\mathbb{L}(\theta_t,\omega_t)\|^2 \\
&\quad + [(1+\alpha)(5\rho^2 l^2 + 2)\frac{L\eta_\theta^2}{2} + \alpha(\frac{l^2\rho^2\eta_\theta}{2} + \frac{l\eta_\theta^2}{2}(5\rho^2 l^2 + 2)) \\
&\quad + \alpha(\frac{l\eta_\omega^2}{2} + \frac{l\eta_\theta^2\eta_\omega}{2}(5\rho^2 l^2 + 2))]\frac{\sigma^2}{M} \qquad (7.26) \\
&\leq -\{\frac{\eta_\theta}{2}(1+\alpha)(1-5\rho l - 2L\eta_\theta(4\rho^2 l^2 + 2\rho l + 2)) \\
&\quad - 2\alpha[(\frac{3\eta_\theta}{2} + \frac{l^2\rho^2\eta_\theta}{2} + l\eta_\theta^2(2\rho^2 l^2 + \rho l + 1)) + l\eta_\theta^2\eta_\omega(2\rho^2 l^2 + \rho l + 1)]\} \\
&\quad \mathbb{E}\|\nabla\mathbb{L}^*(\theta_t)\|^2 + \{(1+\alpha)(\frac{\eta_\theta}{2}(1+\frac{1}{2}\rho l) + L\eta_\theta^2(4\rho^2 l^2 + 2\rho l + 2)) \\
&\quad + 2\alpha[(\frac{3\eta_\theta}{2} + \frac{l^2\rho^2\eta_\theta}{2} + l\eta_\theta^2(2\rho^2 l^2 + \rho l + 1)) + l\eta_\theta^2\eta_\omega(2\rho^2 l^2 + \rho l + 1)]\} \\
&\quad \mathbb{E}\|\nabla\mathbb{L}^*(\theta_t) - \nabla_\theta\mathbb{L}(\theta_t,\omega_t)\|^2 - \alpha(\frac{\eta_\omega}{2} - \frac{l\eta_\omega^2}{2})\mathbb{E}\|\nabla_\omega\mathbb{L}(\theta_t,\omega_t)\|^2 + [(1+\alpha)(5\rho^2 l^2 + 2)\frac{L\eta_\theta^2}{2} \\
&\quad + \alpha(\frac{l^2\rho^2\eta_\theta}{2} + \frac{l\eta_\theta^2}{2}(5\rho^2 l^2 + 2)) + \alpha(\frac{l\eta_\omega^2}{2} + \frac{l\eta_\theta^2\eta_\omega}{2}(5\rho^2 l^2 + 2))]\frac{\sigma^2}{M}. \qquad (7.27)
\end{aligned}
$$

Since we have the following property according to Lemma 2 and the PL condition 12:

$$\|\nabla \mathbb{L}^*(\theta_t) - \nabla_\theta f(\theta_t, \omega_t)\| \leq l\|\omega^*(\theta_t) - \omega_t\| \leq \kappa \|\nabla_\omega f(\theta_t, \omega_t)\|.$$

So we can further the above inequality as follows:

$$\begin{aligned}
&\mathbb{E}[V_{t+1}] - \mathbb{E}[V_t] \\
\leq &-\{\frac{\eta_\theta}{2}(1+\alpha)(1 - 5\rho l - 2L\eta_\theta(4\rho^2 l^2 + 2\rho l + 2)) \\
&- 2\alpha[(\frac{3\eta_\theta}{2} + \frac{l^2\rho^2\eta_\theta}{2} + l\eta_\theta^2(2\rho^2 l^2 + \rho l + 1)) + l\eta_\theta^2\eta_\omega(2\rho^2 l^2 + \rho l + 1)]\}\mathbb{E}\|\nabla \mathbb{L}^*(\theta_t)\|^2 \\
&- \{\alpha(\frac{\eta_\omega}{2} - \frac{l\eta_\omega^2}{2}) - \kappa^2[(1+\alpha)(\frac{\eta_\theta}{2}(1 + \frac{1}{2}\rho l) + L\eta_\theta^2(4\rho^2 l^2 + 2\rho l + 2)) \\
&+ 2\alpha[(\frac{3\eta_\theta}{2} + \frac{l^2\rho^2\eta_\theta}{2} + l\eta_\theta^2(2\rho^2 l^2 + \rho l + 1)) + l\eta_\theta^2\eta_\omega(2\rho^2 l^2 + \rho l + 1)]]\}\mathbb{E}\|\nabla_\omega \mathbb{L}(\theta_t, \omega_t)\|^2 \\
&+ [(1+\alpha)(5\rho^2 l^2 + 2)\frac{L\eta_\theta^2}{2} \\
&+ \alpha(\frac{l^2\rho^2\eta_\theta}{2} + \frac{l\eta_\theta^2}{2}(5\rho^2 l^2 + 2)) + \alpha(\frac{l\eta_\omega^2}{2} + \frac{l\eta_\theta^2\eta_\omega}{2}(5\rho^2 l^2 + 2))]\frac{\sigma^2}{M}.
\end{aligned}$$

Then we require the parameters satisfy: $\alpha = \frac{1}{16}$, $\rho l \leq \frac{1}{16}$, $\eta_\theta(2\rho l + 1)^2 \kappa l \leq \frac{1}{64}$, $\kappa^2\eta_\theta l \leq \frac{1}{128}$, $\rho \leq \frac{\eta_\theta}{2l}$ and $\eta_\omega \leq 64\kappa^2\eta_\theta$.

So the inequality can be further simplified as:

$$\begin{aligned}
&\mathbb{E}[V_{t+1}] - \mathbb{E}[V_t] \\
\leq &-\frac{11}{80}\eta_\theta \mathbb{E}\|\nabla \mathbb{L}^*(\theta_t)\|^2 - \frac{41}{32}\eta_\theta\kappa^2\mathbb{E}\|\nabla_\omega f(\theta_t, \omega_t)\|^2 + 129\kappa^4 l\eta_\theta^2\frac{\sigma^2}{M}.
\end{aligned} \tag{7.28}$$

Telescoping the above inequality we can get:

$$\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}\|\nabla \mathbb{L}^*(\theta_t)\|^2 \leq \frac{80}{11\eta_\theta T}(\mathbb{E}[V_0] - \mathbb{E}[V_T]) + 960\kappa^4 l\eta_\theta\frac{\sigma^2}{M}. \tag{7.29}$$

Further, we can evaluate the first term that:

$$\begin{aligned}\mathbb{E}[V_0] - \mathbb{E}[V_T] &\leq \mathbb{E}[V_0] - \min_{\theta,\omega} \mathbb{E}[V(\theta,\omega)] \\ &\leq \mathbb{E}[\mathbb{L}^*(\theta_0)] - \min_{\theta} \mathbb{E}[\mathbb{L}^*(\theta)] + \frac{1}{16}(\mathbb{E}[\mathbb{L}^*(\theta_0)] - \mathbb{E}[\mathbb{L}(\theta_0,\omega_0)]) \\ &= \mathbb{E}[\mathbb{L}^*(\theta_0)] - \min_{\theta} \mathbb{E}[\mathbb{L}^*(\theta)] + \frac{1}{16}\Delta_0,\end{aligned}$$

where we denote the initial error as: $\Delta_0 = \mathbb{E}[\mathbb{L}^*(\theta_0)] - \mathbb{E}[\mathbb{L}(\theta_0,\omega_0)]$.

Therefore, the inequality 7.29 can be further evaluated as:

$$\frac{1}{T}\sum_{t=0}^{T-1} \mathbb{E}\|\nabla \mathbb{L}^*(\theta_t)\|^2 \leq \frac{80}{11\eta_\theta T}(\mathbb{E}[\mathbb{L}^*(\theta_0)] - \min_{\theta} \mathbb{E}[\mathbb{L}^*(\theta)]) + \frac{5}{11\eta_\theta T}\Delta_0 + 960\kappa^4 l\eta_\theta \frac{\sigma^2}{M}, \tag{7.30}$$

when we select $\eta_\theta = \sqrt{\frac{M(\mathbb{E}[\mathbb{L}^*(\theta_0)] - \min_\theta \mathbb{E}[\mathbb{L}^*(\theta)])}{132T\kappa^4 l\sigma^2}}$, and samples can be minibatch, the convergence can be bounded by:

$$\frac{1}{T}\sum_{t=0}^{T-1} \mathbb{E}\|\nabla \mathbb{L}^*(\theta_t)\|^2 \leq 320\sqrt{\frac{3\kappa^4 l(\mathbb{E}[\mathbb{L}^*(\theta_0)] - \min_\theta \mathbb{E}[\mathbb{L}^*(\theta)])\sigma^2}{11MT}} = \mathcal{O}(\frac{\kappa^2}{\sqrt{MT}}). \tag{7.31}$$

□

7.2 Proof for EVIL

This section provides the proof of Proposition 2. First, by applying the selected invariant parameters, we can have the label prediction $\hat{Y}^e = \mathrm{sgn}(\theta_{inv}^\top Z^e) = \mathrm{sgn}((\mathbf{m} \circ \theta)^\top Z^e)$ where $\mathrm{sgn}(\cdot)$ returns the sign of input value. Further, we have:

$$\begin{aligned}(\mathbf{m} \circ \theta)^\top Z^e &= \frac{1}{\sqrt{M}}\mathbf{m}^\top Z^e &(7.32)\\ &= \sqrt{M}\frac{1}{M}\left[\mathbf{m}_{i\in[0,M_{inv}]}, \mathbf{m}_{i\in[M_{inv},M]}\right]^\top [Z^e_{inv}, Z^e_{var}] \\ &= \sqrt{M}\left[\frac{1}{M_{inv}}\mathbf{m}^\top_{i\in[0,M_{inv}]} Z^e_{inv} + \frac{1}{M_{var}}\mathbf{m}^\top_{i\in[M_{inv},M]} Z^e_{var}\right]. &(7.33)\end{aligned}$$

Then the error produced by the current sparse network for a given environment is:

$$\mathrm{Err}^e = \frac{1}{2}\mathbb{E}_{(X^e,Y^e)\sim e}\left[1 - Y^e\hat{Y}^e\right] = \frac{1}{2}\left[1 - \mathbb{E}^e\left[Y^e\hat{Y}^e\right]\right]. \tag{7.34}$$

Here we simplify the expectation on samples from distribution e as $\mathbb{E}^e$.

$$\begin{aligned}\mathbb{E}^e\left[Y^e\hat{Y}^e\right] &= \mathbb{E}^e\left[\operatorname{sgn}(\frac{1}{M}\mathbf{m}^\top Z^e)Y^e\right] \\ &= \sum_{y\in\{-1,1\}}\mathbb{P}\left[Y^e=y\right]\mathbb{E}^e\left[\operatorname{sgn}(\frac{1}{M}\mathbf{m}^\top Z^e)\mid Y^e=y\right]y. \end{aligned} \tag{7.35}$$

$$\begin{aligned}\mathbb{E}^e\left[\operatorname{sgn}(\frac{1}{M}\mathbf{m}^\top Z^e)\mid Y^e=1\right] &= \mathbb{P}\left[\frac{1}{M}\mathbf{m}^\top Z^e>0\mid Y^e=1\right]-\mathbb{P}\left[\frac{1}{M}\mathbf{m}^\top Z^e\le 0\mid Y^e=1\right] \\ &= 2\mathbb{P}\left[\frac{1}{M}\mathbf{m}^\top Z^e>0\mid Y^e=1\right]-1. \end{aligned} \tag{7.36}$$

Similar to Zhang et al. [184], we observe that:

$$\mathbb{P}\left[\frac{1}{M}\mathbf{m}^\top Z^e\le 0\mid Y^e=1\right]=\mathbb{P}\left[\frac{1}{M}\mathbf{m}^\top Z^e>0\mid Y^e=-1\right], \tag{7.37}$$

and plug Equation (7.36) into Equation (7.35) to get:

$$\mathrm{Err}^e=\mathbb{P}\left[\frac{1}{M}\mathbf{m}^\top Z^e\le 0\mid Y^e=1\right]. \tag{7.38}$$

Similar to Equation (7.33), we further decompose Equation (7.38):

$$\begin{aligned}\mathrm{Err}^e &= \mathbb{P}\left[\frac{1}{M_{inv}}\mathbf{m}^\top_{i\in[0,M_{inv}]}Z^e_{inv}+\frac{1}{M_{var}}\mathbf{m}^\top_{i\in[M_{inv},M]}Z^e_{var}\le 0\mid Y^e=1\right] \\ &\le \mathbb{P}\left[\frac{1}{M_{inv}}\mathbf{m}^\top_{i\in[0,M_{inv}]}Z^e_{inv}\le 0\mid Y^e=1\right]+\mathbb{P}\left[\frac{1}{M_{var}}\mathbf{m}^\top_{i\in[M_{inv},M]}Z^e_{var}\le 0\mid Y^e=1\right]. \end{aligned} \tag{7.39}$$

Since all elements in Z^e_{inv} are equal to Y^e, the first term equals 0, thus the equality also holds. Then, we have:

$$\mathrm{Err}^e=\mathbb{P}\left[\frac{1}{M_{var}}\mathbf{m}^\top_{i\in[M_{inv},M]}Z^e_{var}\le 0\mid Y^e=1\right]. \tag{7.40}$$

Here we assume each element of Z^e and $\mathbf{m}$ are independent with each other, we can have $\mathrm{Err}^e=\mathbb{P}\left[\mathbf{m}_iZ^e_{var,i}\le 0\mid Y^e=1\right]$. It is obvious that there is only one situation when the error occurs, i.e., $\mathbf{m}_i=1$ and $Z^e_{var,i}\le 0$. Only in this scenario, the sparse training strategy would update the mask to value 0. Therefore, $\mathbb{P}\left[\mathbf{m}_i=0\right]=1-p^e$. For other cases where $\mathbb{P}\left[Z^e_{var,i}>0\right]=p^e$, the value of each $\mathbf{m}_i$ is randomly initialized and stays intact since there

is no error, hence, $\mathbb{P}\left[\mathbf{m}_i = 0\right] = \mathbb{P}\left[\mathbf{m}_i = 1\right] = \frac{p^e}{2}$. So, for $i \in [M_{inv}, M]$, $\mathbb{P}\left[\mathbf{m}_i = 0\right] = 1 - \frac{p^e}{2}$ and $\mathbb{P}\left[\mathbf{m}_i = 1\right] = \frac{p^e}{2}$.

To further bound the error, we denote $\overline{[\mathbf{m}Z^e]}_{var} = \frac{1}{M_{var}}\mathbf{m}^\top_{i\in[M_{inv},M]}Z^e_{var}$, and have:

$$\begin{aligned}
\mathrm{Err}^e &= \mathbb{P}\left[\overline{[\mathbf{m}Z^e]}_{var} \leq 0 \mid Y^e = 1\right] \\
&= \mathbb{P}\left[\overline{[\mathbf{m}Z^e]}_{var} - \mathbb{E}\left[\overline{[\mathbf{m}Z^e]}_{var}\right] \leq -\mathbb{E}\left[\overline{[\mathbf{m}Z^e]}_{var}\right] \mid Y^e = 1\right] \\
&\leq \mathbb{P}\left[|\overline{[\mathbf{m}Z^e]}_{var} - \mathbb{E}\left[\overline{[\mathbf{m}Z^e]}_{var}\right]| \geq \mathbb{E}\left[\overline{[\mathbf{m}Z^e]}_{var}\right] \mid Y^e = 1\right]. \qquad (7.41)
\end{aligned}$$

$$\begin{aligned}
\mathbb{E}\left[\overline{[\mathbf{m}Z^e]}_{var}\right] &= \mathbb{E}\left[\mathrm{sgn}(\frac{1}{M_{var}}\mathbf{m}^\top_{i\in[M_{inv},M]}Z^e_{var}) \mid Y^e = 1\right] \\
&= 2\mathbb{P}\left[\frac{1}{M_{var}}\mathbf{m}^\top_{i\in[M_{inv},M]}Z^e_{var} > 0 \mid Y^e = 1\right] - 1 \\
&= (p^e)^2 - 1. \qquad (7.42)
\end{aligned}$$

Therefore, $\mathrm{Err}^e \leq 2e^{-2((p^e)^2-1)^2 M_{var}} = \mathcal{O}(e^{-(p^e)^4})$. In contrast to the idealized bound that achieves 0 error in Zhang et al. [184], when θ is initialized with the unit norm and given a small p^e, the error could be considerably large. This is because the error produced by the variant parameters is largely decided by the probability p^e, which could further affect the pruning process. Based on such an intuition, we propose to enhance the pruning strategy by adding an additional regularization that leverages domain knowledge.

Specifically, the regularization considers the errors from distinguishing different distributions using variant parameters:

$$\begin{aligned}
&\mathrm{Err}^d = \frac{1}{2}\mathbb{E}_{(X,Y)\sim\mathcal{E}}\left[1 - D\hat{D}\right] = \frac{1}{2}\left[1 - \mathbb{E}\left[D\hat{D}\right]\right], \\
&\text{where } \hat{D} = \mathrm{sgn}(\theta^\top_{var}Z^e) = \mathrm{sgn}(((1-\mathbf{m})\circ\theta)^\top Z^e). \qquad (7.43)
\end{aligned}$$

Similar to Equation (7.38), we can have:

$$\begin{aligned}\mathrm{Err}^d &= \mathbb{P}\left[\frac{1}{M}(1-\mathbf{m})^\top Z^e \le 0 \mid D=1\right] \\ &= \mathbb{P}\left[\frac{1}{M_{inv}}(1-\mathbf{m})^\top_{i\in[0,M_{inv}]} Z^e_{inv} + \frac{1}{M_{var}}(1-\mathbf{m})^\top_{i\in[M_{inv},M]} Z^e_{var} \le 0 \mid D=1\right].\end{aligned} \tag{7.44}$$

Since all elements in Z^e_{var} equal to D, we can have:

$$\mathrm{Err}^d = \mathbb{P}\left[\frac{1}{M_{inv}}(1-\mathbf{m})^\top_{i\in[0,M_{inv}]} Z^e_{inv} \le 0 \mid D=1\right]. \tag{7.45}$$

Therefore, for $i \in [0, M_{inv}]$, $\mathbb{P}[\mathbf{m}_i = 1] = 1 - \frac{q^e}{2}$ and $\mathbb{P}[\mathbf{m}_i = 0] = \frac{q^e}{2}$. As a result, the regularization can complement the mask by finding the invariant parameters with at least probability $1 - \frac{q^e}{2}$. Moreover, based on a given sparsity ratio $R = \frac{M_{var}}{M}$, i.e., only M_{inv} elements from Z^e would be selected by $\mathbf{m}$, the erroneous mask that produces classification error can be further constrained from being too much. Particularly, from $\mathbb{P}[\mathbf{m}_i = 1] = 1 - \frac{q^e}{2}, i \in [0, M_{inv}]$, we can have $\mathbb{P}[\mathbf{m}_i = 1] = \frac{M_{inv} - (1-\frac{q^e}{2})M_{inv}}{M_{var}} = \frac{q^e M_{inv}}{2M_{var}}, i \in [M_{inv}, M]$ instead of $\frac{p^e}{2}$ which is calculated based on random initialization. Hence, we can again bound the classification error as:

$$\begin{aligned}\mathbb{E}\left[\overline{[\mathbf{m}Z^e]}_{var}\right] &= 2\mathbb{P}\left[\frac{1}{M_{var}}\mathbf{m}^\top_{i\in[M_{inv},M]} Z^e_{var} > 0 \mid Y^e = 1\right] - 1 \\ &= p^e \frac{q^e M_{inv}}{M_{var}} - 1,\end{aligned} \tag{7.46}$$

$$\begin{aligned}\mathrm{Err}^e &\le \mathbb{P}\left[|\overline{[\mathbf{m}Z^e]}_{var} - \mathbb{E}\left[\overline{[\mathbf{m}Z^e]}_{var}\right]| \ge \mathbb{E}\left[\overline{[\mathbf{m}Z^e]}_{var}\right] \mid Y^e = 1\right] \\ &\le 2e^{-2(\frac{q^e M_{inv}}{M_{var}}p^e - 1)^2 M_{var}} = \mathcal{O}(e^{-(p^e)^2})\end{aligned} \tag{7.47}$$

7.3 Proof for MVT

This section provides the proof of Theorem 5.

Pretraining distribution formulation. We based on the in-context learning framework proposed by Xie et al. [298]. In this framework, a latent concept ϕ from a concept space Φ

defines a pretraining distribution p over prompt tokens o observed from a vocabulary space O. To generate the desired content, we first sample a concept from a prior $p(\phi)$ and then sample the tokens conditioned on the concept. We denote the total length of the pretraining examples is T:

$$p(o_1, \ldots, o_T) = \int_{\phi \in \Phi} p(o_1, \ldots, o_T|\phi) p(\phi) d\phi. \tag{7.48}$$

The conditional probability $p(o_1, \ldots, o_T|\phi)$ is defined by a Hidden Markov Model (HMM). Based on the concept ϕ, a transition matrix of the HMM hidden states $h_1, \ldots, h_T$ from a hidden state space $\mathcal{H}$ can be found.

Prompt distribution formulation. During the in-context learning process, we sample a prompt from a new distribution p_{prompt}, which contains n independent exemplars and 1 query example, which are all conditioned on a shared prompt concept ϕ^*. The goal is to predict the next token based on the exemplars and the query example. Specifically, the i-th exemplar O_i consists of a tokenized image feature $x_i = O_i[1 : k_x]$, a text description to claim the class of the image $y_i = O_i[k_x : k_x + k_y]$, and a binary prediction to judge whether the claim of the image is "True" or "False", which is denoted by $z_i = O_i[k_x + k_y : k_x + k_y + 1]$ at the end of each exemplar. The generating process of the i-th exemplar is as follows:

(1) First generate a start hidden state h_i^{start} from prompt start distribution p_{prompt};
(2) Given h_i^{start}, generate the exemplar sequence $O_i = [x_i, y_i, z_i]$ from $p(O_i|h_i^{start}, \phi^*)$, the generate exemplars are conditioned on a given prompt concept ϕ^*.

The query example Q is sampled similarly without the binary prediction z_q, *i.e.* $Q = [x_q, y_q]$. Between each exemplar and the query example, there is a special delimiter token o^{delim} denoting the end of each exemplar sequence. The prompt can be formulated as follows:

$$[S_n, Q] = [x_1, y_1, z_1, o^{delim}, x_2, y_2, z_2, o^{delim}, \ldots, x_n, y_n, z_n, o^{delim}, x_q, y_q] \sim p_{prompt}, \tag{7.49}$$

where S_n denotes the n independent exemplars for in-context demonstration.

Denoising In-context learning task. In our denoising in-context learning, the output prediction z for each image and text pair $[x, y]$ is sampled based on the prompt distribution

$p_{prompt}(z|x,y)$:

$$z_q \sim p_{prompt}(z|x,y) = \mathbb{E}_{h_q^{start} \sim p_{prompt}(h_q^{start}|Q)}[p(z|Q, h_q^{start}, \phi^*)], \tag{7.50}$$

where h_q^{start} denotes the hidden state corresponding to the first token of Q, *i.e.,* the first token of x_q.

Our goal is to analyze the in-context predictor $f_n(x_q, y_q) = \arg\max_z p(z|S_n, x_q, y_q)$ which outputs the most likely prediction over the pretraining distribution p conditioned on the exemplars S_n sampled from the prompt distribution p_{prompt}. We assume the in-context predictor is trained by $0-1$ error with n training examples:

$$\mathcal{L}_{0-1}(f_n) = \mathbb{E}_{x_q, y_q} \sim p_{prompt}[\mathbf{1}[f_n(x_q) \neq y_q]], \tag{7.51}$$

and same for the prediction z_q:

$$\mathcal{L}_{0-1}(f_n) = \mathbb{E}_{x_q, y_q, z_q} \sim p_{prompt}[\mathbf{1}[f_n(x_q, y_q) \neq z_q]]. \tag{7.52}$$

One major difference of our denoising in-context learning strategy is that we not only use positive exemplars that show exact image-text match, *i.e.,* $(x,y) \sim p(x,y) = p_{prompt}(x,y)$, we also have negative exemplars where image and text are not corresponding to each other. To construct such a prompt, we have to first select the ideal y, based on the matching result of x and y, we can further obtain the prediction z. Therefore, in the following theoretical proof, we propose to conduct two-step analyses on y and z, respectively.

7.3.1 Assumptions

Our theoretical framework is built upon Xie et al. [298], whose assumptions are also applied to our analysis.

ASSUMPTION 16 (Delimiter hidden states). *Let the delimiter hidden states $\mathcal{D}$ be a subset of $\mathcal{H}$. For any $h^{delim} \in \mathcal{D}$ and $\phi \in \Phi$, $p(o^{delim}|h^{delim}, \phi) = 1$ and for any $h \notin \mathcal{D}$, $p(o^{delim}|h, \phi) = 0$.*

ASSUMPTION 17 (Bound on delimiter transitions). *For any delimiter state $h^{delim} \in \mathcal{D}$ and any hidden state $h \in \mathcal{H}$, the probability of transitioning to a delimiter hidden state under ϕ is upper bounded $p(h^{delim}|h, \phi) < c_2$ for any $\phi \in \Phi \setminus \{\phi^*\}$, and is lower bounded $p(h^{delim}|h, \phi^*) > c_1 > 0$ for ϕ^*. Additionally, the start hidden state distribution for delimiter hidden states is bounded as $p(h^{delim}|\phi) \in [c_3, c_4]$.*

ASSUMPTION 18 (Distribution shift from prompt start distribution). *We assume that the prompt start distribution p_{prompt} is close in TV distance to all hidden transition distributions (under ϕ^*) starting from a delimiter hidden state:*

$$\max_{h^{delim} \in \mathcal{D}} TV(p_{prompt}(h) \| p(h|h^{delim}, \phi^*)) < \tau/4. \tag{7.53}$$

Here, $\tau = p_{prompt}(y_{max}|Q) - \max_{y \neq y_{max}} p_{prompt}(y|Q)$ is the margin between the most likely label $y_{max} = \arg\max_y p_{prompt}(y|Q)$ and the second most likely label.

ASSUMPTION 19 (Well-specification). *The prompt concept ϕ^* is in Φ.*

ASSUMPTION 20 (Regularity). *The pretraining distribution p satisfies: (1) Lower bound on transition probability for the prompt concept ϕ^*: for any pair of hidden states $h, h' \in \mathcal{H}$, $p(h|h', \phi^*) > c_5 > 0$. (2) Start hidden state is lower bounded: for any $h \in \mathcal{H}$, $p(h|\phi^*) \geq c_8 > 0$. (3) All tokens can be emitted: for every symbol o, there is some hidden state $h \in \mathcal{H}$ such that $p(o|h, \phi^*) > c_6 > 0$, (4) The prior $p(\phi)$ has support over the entire concept family Φ and is bounded above everywhere.*

Except from the above five adapted assumptions from Xie et al. [298], we have an another mild assumption:

ASSUMPTION 21 (Distribution consistency). *The pretraining distribution p and prompt distribution p_{prompt} satisfy $\forall (x_q, y_q) \sim p_{prompt}, p(x_q, y_q) = p_{prompt}(x_q, y_q)$.*

This assumption indicates that the chosen prompt distribution is a sub-distribution of the pretraining distribution and the joint distribution of x_q and y_q is consistent across p and p_{prompt}. This assumption avoids the situations where there are concept shifts between p and p_{prompt}, *i.e.*, all $y \sim p_{prompt}$ are known in p and can find an exact match for each x_q in p.

7.3.2 Theoretical Proof

We first show that given a query image x_q, when conditioned on a concept ϕ^* and prompt S_n, the most probable text output token for y_q is the same as the class in the prompt distribution p_{prompt} with maximum probability. Then, we show that: in our denoising in-context learning, when achieving the most likely prediction z output by the MLLM predictor, the class text y_q in the pretraining distribution p is the same as the one found by the prompt distribution p_{prompt}, which is the exact match for the give image x_q.

Before we start analyzing the binary prediction z, we first investigate the most probable class y given prompt and query image $\arg\max_y p(y|S_n, x_q)$.

THEOREM 22. *Assume that the above assumptions hold, if for all $\phi \in \Phi$, $\phi \neq \phi^*$, the concept ϕ^* satisfies the distinguishability condition: $\sum_{j=1}^{k} KL_j(\phi^*\|\phi) > \epsilon_{start}^{\phi} + \epsilon_{delim}^{\phi}$, then as $n \to \infty$, the prediction y according to the pretraining distribution is*

$$\arg\max_y p(y|S_n, x_q, \phi^*) \to \arg\max_y p_{prompt}(y|x_q). \tag{7.54}$$

Thus, the in-context predictor f_n achieves the optimal $0-1$ risk: $\lim_{n\to\infty} \mathcal{L}_{0-1}(f_n) = \inf_f \mathcal{L}_{0-1}(f)$.

The detailed proof of this theorem is similar to Xie et al. [298], please refer to the Section D of the original paper.

Under this assumption, the in-context predictor is guaranteed to have the highest probability of generating the class description y that exactly matches the query image x_q. In another way, when x_q does not belong to y, the probability $p(y|S_n, x_q)$ is less than the optimal value.

LEMMA 6. *Under the same condiction of Theorem 22, the prediction z according to the pretraining distribution is*

$$\arg\max_z p(z|S_n, x_q, y_q, \phi^*) \to \arg\max_z p_{prompt}(z|x_q, y_q). \tag{7.55}$$

Lemma 6 can be easily derived based on Theorem 22 by considering y as a fixed prompt.

THEOREM 23. *Assume that the above assumptions hold, as $n \to \infty$, when achieving the largest prediction probability of z given prompt under concept ϕ^*, the corresponding class description y follows the same y obtained from the prompt distribution:*

$$\arg\max_y p(z|S_n, x_q, y, \phi^*) \to \arg\max_y p_{prompt}(z|x_q, y) \tag{7.56}$$

PROOF. Since we already have Theorem 22, if we can prove that

$$\arg\max_y p(y|S_n, x_q, \phi^*) = \arg\max_y p(z|S_n, x_q, y, \phi^*), \tag{7.57}$$

$$\arg\max_y p_{prompt}(y|x_q) = \arg\max_y p_{prompt}(z|x_q, y), \tag{7.58}$$

then we can complete the justification.

$$p(z|S_n, x_q, y, \phi^*) = \sum_{h_q^{start} \in \mathcal{H}} p(z|h_q^{start}) p(h_q^{start}|S_n, x_q, y, \phi^*). \tag{7.59}$$

By expanding the last term, we have:

$$p(h_q^{start}|S_n, x_q, y, \phi^*) = \frac{p(x_q, y|h_q^{start}, S_n, \phi^*) p(h_q^{start})}{p(x_q, y)} \tag{7.60}$$

$$\propto \frac{p(x_q, y|h_q^{start}, S_n, \phi^*)}{p(x_q, y)} \tag{7.61}$$

where $p(h_q^{start})$ is considered as a constant. Moreover, based on Assumption 21, the joint distribution $p(x_q, y)$ is predefined by the pretraining distribution, which does not affect the marginal distribution of z, thus we can have

$$\frac{p(x_q, y|h_q^{start}, S_n, \phi^*)}{p(x_q, y)} = \frac{p(y|S_n, x_q, h_q^{start}, \phi^*) p(x_q|h_q^{start})}{p(x_q, y)} \tag{7.62}$$

$$\propto p(y|S_n, x_q, h_q^{start}, \phi^*) p(x_q|h_q^{start}). \tag{7.63}$$

Since the change of y does not affect the quantity of $p(z|h_q^{start})$, therefore, applying argmax on both sides of the equation holds for finding the optimal y:

$$\arg\max_y p(z|S_n, x_q, y, \phi^*) = \arg\max_y \sum_{h_q^{start} \in \mathcal{H}} p(z|h_q^{start}) p(y|S_n, x_q, h_q^{start}, \phi^*) \tag{7.64}$$

$$= \arg\max_y p(y|S_n, x_q, h_q^{start}, \phi^*). \tag{7.65}$$

Similarly, we have

$$p_{prompt}(z|x_q, y) = \sum_{h_q^{start} \in \mathcal{H}} p(z|h_q^{start}, \phi^*) p_{prompt}(h_q^{start}|x_q, y), \tag{7.66}$$

$$p_{prompt}(h_q^{start}|x_q, y) = \frac{p_{prompt}(x_q, y|h_q^{start}) p_{prompt}(h_q^{start})}{p_{prompt}(x_q, y)} \tag{7.67}$$

$$\propto p_{prompt}(x_q, y|h_q^{start}) \tag{7.68}$$

$$\propto p_{prompt}(y|x_q, h_q^{start}) p_{prompt}(x_q|h_q^{start}), \tag{7.69}$$

$$\arg\max_y p_{prompt}(z|x_q, y) = \arg\max_y \sum_{h_q^{start} \in \mathcal{H}} p_{prompt}(z|h_q^{start}, \phi^*) p_{prompt}(y|x_q, h_q^{start}) \tag{7.70}$$

$$= \arg\max_y p_{prompt}(y|x_q, h_q^{start}), \tag{7.71}$$

where the change of y still does not affect the quantity of $p_{prompt}(z|h_q^{start}, \phi^*)$. Since

$$p(y|S_n, x_q, \phi^*) = \sum_{h_q^{start} \in \mathcal{H}} p(y|h_q^{start}, S_n, x_q, \phi^*) p(h_q^{start}|S_n, x_q, \phi^*), \tag{7.72}$$

$$p_{prompt}(y|x_q) = \sum_{h_q^{start} \in \mathcal{H}} p_{prompt}(y|h_q^{start}, x_q) p_{prompt}(h_q^{start}|x_q), \tag{7.73}$$

it is easy to find that

$$\arg\max_y p_{prompt}(y|x_q, h_q^{start}) = \arg\max_y p_{prompt}(y|x_q), \tag{7.74}$$

$$\arg\max_y p(y|S_n, x_q, h_q^{start}, \phi^*) = \arg\max_y p(y|S_n, x_q, \phi^*). \tag{7.75}$$

Thus, we have that as $n \to \infty$,

$$\arg\max_y p(z|S_n, x_q, y, \phi^*) \to \arg\max_y p_{prompt}(z|x_q, y). \tag{7.76}$$

□

Lemma 6 and Theorem 23 together show that when given a query image x_q, if the chosen query class description y_q is the true class of x_q, then under the given assumptions, the binary prediction z for judging the correctness of the image-text pair would be the maximum value compared to all other class descriptions $y \neq y_q, y \in \mathcal{Y}$. Therefore, we can justify that using an in-context predictor can help identify the true class label of a given image.

7.4 Proof for OOM

This section provides the proof of Theorem 7.

7.4.1 Lower Bound of Our VIB framework

Recall that we have the following factorization:

$$p(X^{\mathrm{I}}, X^{\mathrm{O}}, V, Y) = p(V, Y|X^{\mathrm{O}}, X^{\mathrm{I}})p(X^{\mathrm{O}}|X^{\mathrm{I}})P(X^{\mathrm{I}}), \tag{7.77}$$

with Markov chains $V \leftrightarrow X^{\mathrm{I}} \leftrightarrow X^{\mathrm{O}}$ and $X^{\mathrm{I}} \leftrightarrow Y \not\leftrightarrow X^{\mathrm{O}}$. Our goal is to maximize the information redundancy [331, 342]:

$$\max I(X^{\mathrm{O}}; X^{\mathrm{I}}; Y) = I(X^{\mathrm{O}}; X^{\mathrm{I}}) - I(X^{\mathrm{O}}; X^{\mathrm{I}}|Y), \tag{7.78}$$

where the first term is lower-bounded by:

$$I(X^{\mathrm{O}}; X^{\mathrm{I}}) \geq \int dx^{\mathrm{O}} dx^{\mathrm{I}} dv p(x^{\mathrm{I}}) p(x^{\mathrm{O}}|x^{\mathrm{I}}) p(v|x^{\mathrm{I}}) \log q(x^{\mathrm{O}}|v) p(v|x^{\mathrm{I}}), \tag{7.79}$$

Then, we consider the second term $I(X^{\mathrm{O}}; X^{\mathrm{I}}|Y)$:

$$I(X^{\mathrm{O}}; X^{\mathrm{I}}|Y) = \int dx^{\mathrm{O}} dx^{\mathrm{I}} dy p(x^{\mathrm{O}}, x^{\mathrm{I}}, y) \log \frac{p(x^{\mathrm{O}}, x^{\mathrm{I}}|y)}{p(x^{\mathrm{O}}|y)p(x^{\mathrm{I}}|y)} \tag{7.80}$$

$$= \int dx^{\mathrm{O}} dx^{\mathrm{I}} dy p(x^{\mathrm{O}}, x^{\mathrm{I}}, y) \log \frac{p(x^{\mathrm{O}}, x^{\mathrm{I}}, y)}{p(y|x^{\mathrm{O}})} - H(Y) + H(Y|X^{\mathrm{I}}) + H(X^{\mathrm{O}}) + H(X^{\mathrm{I}}). \tag{7.81}$$

Note that we use the factorization $p(x^{\mathrm{O}}, x^{\mathrm{I}}, y) = p(y|x^{\mathrm{I}})p(x^{\mathrm{O}}|x^{\mathrm{I}})p(x^{\mathrm{I}})$, and further ignore the entropy terms[1], then we have:

$$I(X^{\mathrm{O}}; X^{\mathrm{I}}|Y) \leq \int dx^{\mathrm{O}} dx^{\mathrm{I}} dy p(y|x^{\mathrm{I}})p(x^{\mathrm{O}}|x^{\mathrm{I}})p(x^{\mathrm{I}}) \log p(y|x^{\mathrm{I}})p(x^{\mathrm{O}}|x^{\mathrm{I}})p(x^{\mathrm{I}}) - \log h(y|x^{\mathrm{O}}), \tag{7.82}$$

which is based on the positivity of KL divergence between our classifier $h(y|x^{\mathrm{O}})$ and $p(y|x^{\mathrm{O}})$.

To this end, we can lower-bound our objective by combining Equations (7.79) and (7.82):

$$I(X^{\mathrm{O}}; X^{\mathrm{I}}; Y) \geq \int dx^{\mathrm{O}} dx^{\mathrm{I}} dv p(x^{\mathrm{I}})p(x^{\mathrm{O}}|x^{\mathrm{I}})p(v|x^{\mathrm{I}}) \log q(x^{\mathrm{O}}|v)p(v|x^{\mathrm{I}}) \tag{7.83}$$

$$- \int dx^{\mathrm{O}} dx^{\mathrm{I}} dy p(y|x^{\mathrm{I}})p(x^{\mathrm{O}}|x^{\mathrm{I}})p(x^{\mathrm{I}}) \log p(y|x^{\mathrm{I}})p(x^{\mathrm{O}}|x^{\mathrm{I}})p(x^{\mathrm{I}}) + \log h(y|x^{\mathrm{O}}) = \mathcal{L}_{con}. \tag{7.84}$$

7.4.2 Proof of Theorem

Now we start the proof of Theorem 7.

PROOF.

ASSUMPTION 24 (Relaxed triangle inequality). *For the distance function $d : \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}^+$, there exists $c_d \geq 1$ such that $\forall \hat{y}_1, \hat{y}_2, \hat{y}_3 \in \hat{\mathcal{Y}} d(\hat{y}_1, \hat{y}_2) \leq c_d(d(\hat{y}_1, \hat{y}_3) + d(\hat{y}_2, \hat{y}_3))$.*

ASSUMPTION 25 (Inverse Lipschitz condition). *For the function d, it holds that $\forall h$,*

$$\mathbb{E}[d(h(x_1, x_2), h^*(x_1, x_2))] \leq |\mathcal{L}(h) - \mathcal{L}(h^*)|, \tag{7.85}$$

where h^* is the Bayes optimal classifier on both x_1 and x_2; and

$$\mathbb{E}[d(h(x), h^*(x))] \leq |\mathcal{L}(h) - \mathcal{L}(h^*)|, \tag{7.86}$$

where h^* is the Bayes optimal classifier on x.

ASSUMPTION 26 (Classifier optimality). *For any classifiers h in comparison to the Bayes' optimal classifier h^*, there exists constants $\epsilon > 0$ such that $|\mathcal{L}(h) - \mathcal{L}(h^*)|^2 \leq \epsilon$.*

[1]We focus on the optimization of $p(Y|X^{\mathrm{O}})$, and $p(Y|X^{\mathrm{I}})$ is given and frozen in our setting.

To bridge h_1^* and h_2^*, we use $h_{1,2}^*$ and h^* to denote the Bayes' optimal classifier on both IM data and all data, respectively. Then, we capture the relationship between the uniqueness of OOM data given both IM data and the difference in their Bayes' optimal prediction errors:

$$\begin{aligned}
|\mathcal{L}(h_{1,2}^*) - \mathcal{L}(h^*)|^2 &= |\mathbb{E}_X \mathbb{E}_{Y|X_1^{\mathrm{I}},X_2^{\mathrm{I}},X^{\mathrm{O}}} \ell(h^*(x_1^{\mathrm{I}}, x_2^{\mathrm{I}}, x^{\mathrm{O}}), y) \\
&\quad - \mathbb{E}_{X_1^{\mathrm{I}},X_2^{\mathrm{I}}} \mathbb{E}_{Y|X_1^{\mathrm{I}},X_2^{\mathrm{I}}} \ell(h_1^*(x_1^{\mathrm{I}}, X_2^{\mathrm{I}}), y)|^2 && (7.87) \\
&\leq |\mathbb{E}_{Y|X_1^{\mathrm{I}},X_2^{\mathrm{I}},X^{\mathrm{O}}} \ell(h^*(x_1^{\mathrm{I}}, x_2^{\mathrm{I}}, x^{\mathrm{O}}), y) - \mathbb{E}_{Y|X_1^{\mathrm{I}},X_2^{\mathrm{I}}} \ell(h_1^*(x_1^{\mathrm{I}}, X_2^{\mathrm{I}}), y)|^2 && (7.88) \\
&\leq \mathrm{KL}(p(y|x_1^{\mathrm{I}}, x_2^{\mathrm{I}}, x^{\mathrm{O}}) \parallel p(y|x_1^{\mathrm{I}}, x_2^{\mathrm{I}})) && (7.89) \\
&\leq \mathbb{E}_X \mathrm{KL}(p(y|x_1^{\mathrm{I}}, x_2^{\mathrm{I}}, x^{\mathrm{O}}) \parallel p(y|x_1^{\mathrm{I}}, x_2^{\mathrm{I}})) && (7.90) \\
&= I(X^{\mathrm{O}}, Y|X_1^{\mathrm{I}}, X_2^{\mathrm{I}}). && (7.91)
\end{aligned}$$

Then, we first capture the redundancy between one IM and OOM data given another IM data:

$$\begin{aligned}
|\mathcal{L}(h_1^*) - \mathcal{L}(h^*)|^2 &= |\mathbb{E}_X \mathbb{E}_{Y|X_1^{\mathrm{I}},X_2^{\mathrm{I}},X^{\mathrm{O}}} \ell(h^*(x_1^{\mathrm{I}}, x_2^{\mathrm{I}}, x^{\mathrm{O}}), y) - \mathbb{E}_{X_1^{\mathrm{I}}} \mathbb{E}_{Y|X_1^{\mathrm{I}}} \ell(h_1^*(x_1^{\mathrm{I}}), y)|^2 && (7.92) \\
&\leq |\mathbb{E}_{Y|X_1^{\mathrm{I}},X_2^{\mathrm{I}},X^{\mathrm{O}}} \ell(h^*(x_1^{\mathrm{I}}, x_2^{\mathrm{I}}, x^{\mathrm{O}}), y) - \mathbb{E}_{Y|X_1^{\mathrm{I}}} \ell(h_1^*(x_1^{\mathrm{I}}), y)|^2 && (7.93) \\
&\leq \mathrm{KL}(p(y|x_1^{\mathrm{I}}, x_2^{\mathrm{I}}, x^{\mathrm{O}}) \parallel p(y|x_1^{\mathrm{I}})) && (7.94) \\
&\leq \mathbb{E}_X \mathrm{KL}(p(y|x_1^{\mathrm{I}}, x_2^{\mathrm{I}}, x^{\mathrm{O}}) \parallel p(y|x_1^{\mathrm{I}})) && (7.95) \\
&= I(X^{\mathrm{O}}, X_2^{\mathrm{I}}, Y|X_1^{\mathrm{I}}). && (7.96)
\end{aligned}$$

Further leveraging triangle inequality through the Bayes' optimal classifier h^* and the inverse Lipschitz condition, we have:

$$\begin{aligned}
\mathbb{E}_{p(x_1^{\mathrm{I}},x_2^{\mathrm{I}},x^{\mathrm{O}})}[d(h_1^*, h_{1,2}^*)] &\leq \mathbb{E}_{p(x_1^{\mathrm{I}},x_2^{\mathrm{I}},x^{\mathrm{O}})}[d(h_1^*, h^*)] + \mathbb{E}_{p(x_1^{\mathrm{I}},x_2^{\mathrm{I}},x^{\mathrm{O}})}[d(h^*, h_{1,2}^*)] && (7.97) \\
&\leq |\mathcal{L}(h_1^*) - \mathcal{L}(h^*)|^2 + |\mathcal{L}(h^*) - \mathcal{L}(h_{1,2}^*)|^2 && (7.98) \\
&\leq I(X^{\mathrm{O}}, X_2^{\mathrm{I}}, Y|X_1^{\mathrm{I}}) + I(X^{\mathrm{O}}, Y|X_1^{\mathrm{I}}, X_2^{\mathrm{I}}). && (7.99)
\end{aligned}$$

Symmetrically, we can have $|\mathcal{L}(h_2^*) - \mathcal{L}(h^*)|^2 \leq I(X^{\mathrm{O}}, X_1^{\mathrm{I}}, Y|X_2^{\mathrm{I}})$ and further obtain:

$$\mathbb{E}_{p(x_2^{\mathrm{I}},x_2^{\mathrm{I}},x^{\mathrm{O}})}[d(h_2^*, h_{1,2}^*)] \leq I(X^{\mathrm{O}}, X_1^{\mathrm{I}}, Y|X_2^{\mathrm{I}}) + I(X^{\mathrm{O}}, Y|X_1^{\mathrm{I}}, X_2^{\mathrm{I}}). \tag{7.100}$$

Then combining with Equation (7.99):

$$\mathbb{E}_{p(x_1^{\mathrm{I}},x_2^{\mathrm{I}})}[d(h_1^*,h_2^*)] \leq I(X^{\mathrm{O}},X_2^{\mathrm{I}},Y|X_1^{\mathrm{I}}) + I(X^{\mathrm{O}},X_1^{\mathrm{I}},Y|X_2^{\mathrm{I}}) + 2I(X^{\mathrm{O}},Y|X_1^{\mathrm{I}},X_2^{\mathrm{I}}) \quad (7.101)$$

Finally, based on the decomposition of the task-related mutual information of X^{O}:

$$I(X^{\mathrm{O}},Y) = I(X^{\mathrm{O}},X_2^{\mathrm{I}},Y|X_1^{\mathrm{I}}) + I(X^{\mathrm{O}},X_1^{\mathrm{I}},Y|X_2^{\mathrm{I}}) + I(X^{\mathrm{O}},Y|X_1^{\mathrm{I}},X_2^{\mathrm{I}}) + I(X^{\mathrm{O}},X_1^{\mathrm{I}},X_2^{\mathrm{I}},Y), \quad (7.102)$$

as shown in Figure 6.3, we can achieve:

$$\alpha(h_1^*,h_2^*) := \mathbb{E}_{p(x_1^{\mathrm{I}},x_2^{\mathrm{I}})}[d(h_1^*,h_2^*)] \leq I(X^{\mathrm{O}},Y) - I(X^{\mathrm{O}},X_1^{\mathrm{I}},X_2^{\mathrm{I}},Y) + I(X^{\mathrm{O}},Y|X_1^{\mathrm{I}},X_2^{\mathrm{I}}), \quad (7.103)$$

□

CHAPTER 8

Conclusion

In this thesis, we provide a systematic study on Trustworthy Machine Learning under Distribution Shifts by exploring capability and responsibility under various practical applications.

We first investigate perturbation shifts in Chapters 2 and 3, where both the type and strength of distribution shift are studied. Specifically, different type refers to benigh OOD data with changed style and malign OOD data with unseen content; and strengths denotes the intensity of corruption applyed to the OOD data. To solve these problems, we proposed a causal framework to understand the data generation process and a robust optimization strategy to effectively enhance the generalization performance against different types of shifts with varied strengths, demonstrating both explainability, robustness, and adaptability for OOD data.

Then, we study domain shifts in Chapter 3, Chapter 4, and Chapter 5. Particularly, we develop multiple optimization strategies to handle OOD data from various domains, namely, worst-case optimization, sparse training, and machine-supervised training. For all these novel techniques, we aim to identify the critical knowledge, i.e., the invariant features in the data. In this way, our approaches can effectively generalize, meanwhile demonstrating great adaptability across domains.

Finally, we look into a novel distribution shift, modality shift, and study it in Chapters 5 and 6. We introduce this problem by revealing the modality gap between LLMs and vision models, which can be addressed by in-context learning to provide learnable signals across modalities. Further, we extend this problem to uncommon modalities and propose OOM

generalization to uncover the hidden knowledge in novel modalities. We provide an information theoretic analysis to extract sharable knowledge, which enables generalize across modalities effectively.

This thesis inspires several extendable studies in the future. First of all, how to achieve general intelligence with alignment with human values is one of the most critical questions to be answered. We are in a pivotal period where there is a rise in general intelligence. When AI is taking over various tasks from different dimensions, i.e., visual tasks, language tasks, and physical tasks. It is urgent the develop a systematic, trustworthy protocol to uniformly regulate every possible dimension. Further, we observe various complexities of different distribution shifts; existing AI models lack universal capability, hindering their general practice. For example, LLMs are still suffering from slight perturbation or adversarial attacks. Therefore, future AIs are required to possess reliable capabilities across various shifts. Finally, AIs are driven by data. However, for some tasks, e.g., modality shift, rare modalities significantly lack data to generalize. Thus, it is critical to develop self-reflective and trustworthy AIs that can discover rules and patterns to explore an unknown knowledge scope.

Bibliography

[1] Yann LeCun, Yoshua Bengio and Geoffrey Hinton. 'Deep learning'. In: *nature* 521.7553 (2015), pp. 436–444.

[2] Alex Krizhevsky, Ilya Sutskever and Geoffrey E Hinton. 'Imagenet classification with deep convolutional neural networks'. In: *Advances in neural information processing systems*. Vol. 25. Curran Associates, Inc., 2012.

[3] Kaiming He et al. 'Deep residual learning for image recognition'. In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2016, pp. 770–778.

[4] Gao Huang et al. 'Densely connected convolutional networks'. In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2017, pp. 4700–4708.

[5] Ashish Vaswani et al. 'Attention is all you need'. In: *Advances in neural information processing systems*. Vol. 30. Curran Associates, Inc., 2017.

[6] Jacob Devlin et al. 'BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding'. In: *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. Vol. 1. Association for Computational Linguistics, 2019, pp. 4171–4186.

[7] Alexey Dosovitskiy et al. 'An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale'. In: *International Conference on Learning Representations (ICLR)*. 2021.

[8] Jared Kaplan et al. 'Scaling Laws for Neural Language Models'. In: *arXiv preprint arXiv:2001.08361* (2020). URL: https://arxiv.org/abs/2001.08361.

[9] Tom B. Brown et al. 'Language Models are Few-Shot Learners'. In: *Advances in Neural Information Processing Systems (NeurIPS)* 33 (2020).

[10] Meta AI. *The Llama 3 Herd of Models*. Tech. rep. arXiv:2407.21783. Meta AI, 2024. URL: https://arxiv.org/abs/2407.21783.

[11] Jinze Bai et al. *Qwen Technical Report*. arXiv preprint. arXiv:2309.16609. 2023. URL: https://arxiv.org/abs/2309.16609.

[12] OpenAI. *ChatGPT: Optimizing Language Models for Dialogue*. Accessed: 2025-12-14. 2022. URL: https://openai.com/blog/chatgpt/.

[13] Gemini Team et al. 'Gemini: A Family of Highly Capable Multimodal Models'. In: *arXiv preprint arXiv:2312.11805* (2023). URL: https://arxiv.org/abs/2312.11805.

[14] Anthropic. *The Claude 3 Model Family: Opus, Sonnet, Haiku*. Technical Report. 2024. URL: https://www-cdn.anthropic.com.

[15] DeepSeek-AI. *DeepSeek LLM: Scaling Open-Source Language Models with Long Context and Strong Reasoning*. arXiv preprint. arXiv:2401.02954. 2024. URL: https://arxiv.org/abs/2401.02954.

[16] Luc Devroye, László Györfi and Gábor Lugosi. *A Probabilistic Theory of Pattern Recognition*. Springer, 1996.

[17] Vladimir N. Vapnik. *Statistical Learning Theory*. New York: Wiley, 1998.

[18] Vladimir Vapnik and Alexey Chervonenkis. 'A note on one class of perceptrons'. In: *Automation and Remote Control* 25 (1963), pp. 821–837.

[19] David E Rumelhart, Geoffrey E Hinton and Ronald J Williams. 'Learning representations by back-propagating errors'. In: *nature* 323.6088 (1986), pp. 533–536.

[20] Kush R Vashney. *Trustworthy machine learning*. Independently published, 2022.

[21] Paul F Christiano et al. 'Deep reinforcement learning from human preferences'. In: *Advances in neural information processing systems* 30 (2017).

[22] Jan Betley et al. 'Emergent Misalignment: Narrow finetuning can produce broadly misaligned LLMs'. In: *arXiv preprint arXiv:2502.17424* (2025).

[23] Preethi Lahoti et al. 'Fairness without demographics through adversarially reweighted learning'. In: *Advances in neural information processing systems* 33 (2020), pp. 728–740.

[24] Lucas Bourtoule et al. ‘Machine unlearning’. In: *2021 IEEE symposium on security and privacy (SP)*. IEEE. 2021, pp. 141–159.

[25] Masashi Sugiyama, Matthias Krauledat and Klaus-Robert Müller. ‘Covariate shift adaptation by importance weighted cross validation.’ In: *Journal of Machine Learning Research* 8.5 (2007).

[26] Kun Zhang et al. ‘Domain adaptation under target and conditional shift’. In: *International conference on machine learning*. Pmlr. 2013, pp. 819–827.

[27] Peter Vorburger and Abraham Bernstein. ‘Entropy-based concept shift detection’. In: *Sixth International Conference on Data Mining (ICDM’06)*. IEEE. 2006, pp. 1113–1118.

[28] Dan Hendrycks and Kevin Gimpel. ‘A baseline for detecting misclassified and out-of-distribution examples in neural networks’. In: *ICLR*. 2017.

[29] Sinno Jialin Pan and Qiang Yang. ‘A survey on transfer learning’. In: *IEEE Transactions on knowledge and data engineering (TKDE)* 22.10 (2009), pp. 1345–1359.

[30] Yaroslav Ganin et al. ‘Domain-adversarial training of neural networks’. In: *Journal of Machine Learning Research (JMLR)* 17.1 (2016), pp. 2096–2030.

[31] Eric Tzeng et al. ‘Adversarial discriminative domain adaptation’. In: *CVPR*. 2017, pp. 7167–7176.

[32] Walter J Scheirer et al. ‘Toward open set recognition’. In: *IEEE transactions on pattern analysis and machine intelligence (TPAMI)* 35.7 (2012), pp. 1757–1772.

[33] Matthias Hein, Maksym Andriushchenko and Julian Bitterwolf. ‘Why relu networks yield high-confidence predictions far away from the training data and how to mitigate the problem’. In: *CVPR*. 2019, pp. 41–50.

[34] Kimin Lee et al. ‘A simple unified framework for detecting out-of-distribution samples and adversarial attacks’. In: *NeurIPS*. Vol. 31. 2018.

[35] Shiyu Liang, Yixuan Li and Rayadurgam Srikant. ‘Enhancing the reliability of out-of-distribution image detection in neural networks’. In: *ICLR*. 2018.

[36] Weitang Liu et al. ‘Energy-based out-of-distribution detection’. In: *NeurIPS*. Vol. 33. 2020, pp. 21464–21475.

[37] Haotao Wang et al. ‘Partial and Asymmetric Contrastive Learning for Out-of-Distribution Detection in Long-Tailed Recognition’. In: *ICML*. PMLR. 2022, pp. 23446–23458.

[38] Qizhou Wang et al. ‘Out-of-distribution Detection with Implicit Outlier Transformation’. In: *ICLR*. 2023.

[39] Qizhou Wang et al. ‘Watermarking for Out-of-distribution Detection’. In: *NeurIPS*. 2022.

[40] Hong Liu et al. ‘Separate to adapt: Open set domain adaptation via progressive separation’. In: *CVPR*. 2019, pp. 2927–2936.

[41] Kuniaki Saito et al. ‘Open set domain adaptation by backpropagation’. In: *ECCV*. 2018, pp. 153–168.

[42] Zhuo Huang et al. ‘Universal Semi-Supervised Learning’. In: *NeurIPS*. Vol. 34. 2021.

[43] Zhuo Huang, Jian Yang and Chen Gong. ‘They are not completely useless: Towards recycling transferable unlabeled data for class-mismatched semi-supervised learning’. In: *IEEE Transactions on Multimedia* (2022).

[44] Zhuo Huang et al. ‘Flatmatch: Bridging labeled data and unlabeled data with cross-sharpness for semi-supervised learning’. In: *NeurIPS*. Vol. 36. 2023, pp. 18474–18494.

[45] Zhe Huang et al. ‘Fix-A-Step: Effective Semi-supervised Learning from Uncurated Unlabeled Sets’. In: *arXiv preprint arXiv:2208.11870* (2022).

[46] Mingyu Li et al. ‘Dynamic Weighted Adversarial Learning for Semi-Supervised Classification under Intersectional Class Mismatch’. In: *ACM Transactions on Multimedia Computing, Communications and Applications* 20.4 (2024), pp. 1–24.

[47] Avital Oliver et al. ‘Realistic evaluation of deep semi-supervised learning algorithms’. In: *NeurIPS*. 2018, pp. 3235–3246.

[48] Kuniaki Saito, Donghyun Kim and Kate Saenko. ‘OpenMatch: Open-Set Semi-supervised Learning with Open-set Consistency Regularization’. In: *NeurIPS*. Vol. 34. 2021.

[49] Qing Yu et al. ‘Multi-task curriculum framework for open-set semi-supervised learning’. In: *ECCV*. 2020.

[50] Yoshua Bengio et al. ‘Deep learners benefit more from out-of-distribution examples’. In: *AISTATS*. JMLR Workshop and Conference Proceedings. 2011, pp. 164–172.

[51] Yaroslav Ganin and Victor Lempitsky. 'Unsupervised domain adaptation by backpropagation'. In: *ICML*. PMLR. 2015, pp. 1180–1189.

[52] Ekin D Cubuk et al. 'Autoaugment: Learning augmentation policies from data'. In: *CVPR*. 2018.

[53] Qizhe Xie et al. 'Unsupervised data augmentation for consistency training'. In: *NeurIPS*. 2020.

[54] David Berthelot et al. 'Mixmatch: A holistic approach to semi-supervised learning'. In: *NeurIPS*. 2019, pp. 5049–5059.

[55] Kihyuk Sohn et al. 'Fixmatch: Simplifying semi-supervised learning with consistency and confidence'. In: *arXiv preprint arXiv:2001.07685* (2020).

[56] Hongxin Wei et al. 'Mitigating neural network overconfidence with logit normalization'. In: *ICML*. PMLR. 2022, pp. 23631–23644.

[57] Hongxin Wei et al. 'Open-Sampling: Exploring Out-of-Distribution data for Rebalancing Long-tailed datasets'. In: *ICLR*. PMLR. 2022, pp. 23615–23630.

[58] Shu Kong and Deva Ramanan. 'Opengan: Open-set recognition via open data generation'. In: *ICCV*. 2021, pp. 813–822.

[59] Abhishek Sinha et al. 'Negative Data Augmentation'. In: *ICLR*. 2020.

[60] Madelyn Glymour, Judea Pearl and Nicholas P Jewell. *Causal inference in statistics: A primer*. John Wiley & Sons, 2016.

[61] Judea Pearl. *Causality*. Cambridge university press, 2009.

[62] Erdun Gao et al. 'MissDAG: Causal Discovery in the Presence of Missing Data with Continuous Additive Noise Models'. In: *NeurIPS*. 2022.

[63] David M Blei, Alp Kucukelbir and Jon D McAuliffe. 'Variational inference: A review for statisticians'. In: *Journal of the American Statistical Association (JASA)* 112.518 (2017), pp. 859–877.

[64] Yu Yao et al. 'Instance-dependent label-noise learning under a structural causal model'. In: *NeurIPS*. Vol. 34. 2021, pp. 4409–4420.

[65] Xiaobo Xia et al. 'Pluralistic Image Completion with Gaussian Mixture Models'. In: *NeurIPS*. 2022.

[66] Katherine Hermann, Ting Chen and Simon Kornblith. 'The origins and prevalence of texture bias in convolutional neural networks'. In: *NeurIPS*. Vol. 33. 2020, pp. 19000–19015.

[67] Yingwei Li et al. 'Shape-texture debiased neural network training'. In: *ICLR*. 2021.

[68] Yonggang Zhang et al. 'CausalAdv: Adversarial Robustness through the Lens of Causality'. In: *ICLR*. 2022.

[69] Martin Arjovsky et al. 'Invariant risk minimization'. In: *arXiv preprint arXiv:1907.02893* (2019).

[70] Lan-Zhe Guo et al. 'Safe deep semi-supervised learning for unseen-class unlabeled data'. In: *ICML*. 2020.

[71] Yanbei Chen et al. 'Semi-supervised learning under class distribution mismatch.' In: *AAAI*. 2020.

[72] Rundong He et al. 'Safe-Student for Safe Deep Semi-Supervised Learning with Unseen-Class Unlabeled Data'. In: *CVPR*. 2022, pp. 14585–14594.

[73] Rundong He et al. 'Not all parameters should be treated equally: Deep safe semi-supervised learning under class distribution mismatch'. In: *AAAI*. Vol. 36. 6. 2022, pp. 6874–6883.

[74] Jie Ren et al. 'Likelihood ratios for out-of-distribution detection'. In: *NeurIPS*. Vol. 32. 2019.

[75] Xiaobo Xia et al. 'Extended T: Learning with Mixed Closed-set and Open-set Noisy Labels'. In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* (2022).

[76] Lawrence Neal et al. 'Open set learning with counterfactual images'. In: *ECCV*. 2018, pp. 613–628.

[77] Xiaobo Xia et al. 'Instance correction for learning with open-set noisy labels'. In: *arXiv preprint arXiv:2106.00455* (2021).

[78] ZongYuan Ge et al. 'Generative openmax for multi-class open set classification'. In: *arXiv preprint arXiv:1707.07418* (2017).

[79] Shreyas Padhy et al. 'Revisiting one-vs-all classifiers for predictive uncertainty and out-of-distribution detection in neural networks'. In: *arXiv preprint arXiv:2007.05134* (2020).

[80] Kaidi Cao, Maria Brbic and Jure Leskovec. 'Open-World Semi-Supervised Learning'. In: *ICLR*. 2022. URL: https://openreview.net/forum?id=O-r8LOR-CCA.

[81] Guangrui Li et al. 'Domain consensus clustering for universal domain adaptation'. In: *CVPR*. 2021, pp. 9757–9766.

[82] Kuniaki Saito et al. 'Universal domain adaptation through self supervision'. In: *NeurIPS*. 2020.

[83] Diederik P Kingma and Max Welling. 'Auto-encoding variational bayes'. In: *arXiv preprint arXiv:1312.6114* (2013).

[84] Mingming Gong et al. 'Domain adaptation with conditional transferable components'. In: *ICML*. PMLR. 2016, pp. 2839–2848.

[85] Bernhard Schölkopf et al. 'Robust learning via cause-effect models'. In: *arXiv preprint arXiv:1112.2738* (2011).

[86] Haoliang Li et al. 'Domain generalization with adversarial feature learning'. In: *CVPR*. 2018, pp. 5400–5409.

[87] Shiv Shankar et al. 'Generalizing across domains via cross-gradient training'. In: *arXiv preprint arXiv:1804.10745* (2018).

[88] Mohammad Taha Bahadori et al. 'Causal regularization'. In: *arXiv preprint arXiv:1702.02604* (2017).

[89] Zheyan Shen et al. 'Causally regularized learning with agnostic data selection bias'. In: *ACM Multimedia*. 2018, pp. 411–419.

[90] Jungsoo Lee et al. 'Learning debiased representation via disentangled feature augmentation'. In: *NeurIPS*. Vol. 34. 2021.

[91] Junhyun Nam et al. 'Learning from failure: De-biasing classifier from biased classifier'. In: *NeurIPS*. Vol. 33. 2020, pp. 20673–20684.

[92] Maximilian Ilse, Jakub M Tomczak and Patrick Forré. 'Selecting data augmentation for simulating interventions'. In: *International Conference on Machine Learning*. PMLR. 2021, pp. 4555–4562.

[93] Jovana Mitrovic et al. 'Representation learning via invariant causal mechanisms'. In: *arXiv preprint arXiv:2010.07922* (2020).

[94] Julius Von Kügelgen et al. 'Self-supervised learning with data augmentations provably isolates content from style'. In: *NeurIPS*. Vol. 34. 2021.

[95] Dong-Hyun Lee. 'Pseudo-label: The simple and efficient semi-supervised learning method for deep neural networks'. In: *ICML Workshop*. 2013.

[96] Ian Goodfellow et al. 'Generative adversarial nets'. In: *NeurIPS*. 2014, pp. 2672–2680.

[97] Takeru Miyato et al. 'Virtual adversarial training: a regularization method for supervised and semi-supervised learning'. In: *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)* 41.8 (2018), pp. 1979–1993.

[98] Riccardo Volpi et al. 'Generalizing to unseen domains via adversarial data augmentation'. In: *NeurIPS*. Vol. 31. 2018.

[99] Aleksander Madry et al. 'Towards deep learning models resistant to adversarial attacks'. In: *ICLR*. 2018.

[100] Qizhou Wang et al. 'Probabilistic margins for instance reweighting in adversarial training'. In: *NeurIPS*. Vol. 34. 2021, pp. 23258–23269.

[101] Sergey Zagoruyko and Nikos Komodakis. 'Wide residual networks'. In: *BMVC*. 2016.

[102] Kaichao You et al. 'Universal Domain Adaptation'. In: *CVPR*. 2020.

[103] Zhangjie Cao et al. 'Learning to transfer examples for partial domain adaptation'. In: *CVPR*. 2019, pp. 2985–2994.

[104] Olga Russakovsky et al. 'Imagenet large scale visual recognition challenge'. In: *International journal of computer vision* 115.3 (2015), pp. 211–252.

[105] Ekin Dogus Cubuk et al. 'RandAugment: Practical Automated Data Augmentation with a Reduced Search Space'. In: *NeurIPS*. Vol. 33. 2020, pp. 18613–18624.

[106] Yuval Netzer et al. 'Reading digits in natural images with unsupervised feature learning'. In: *NeurIPS Workshop*. 2011.

[107] Alex Krizhevsky, Geoffrey Hinton et al. 'Learning multiple layers of features from tiny images'. In: (2009).

[108] Fisher Yu et al. 'Lsun: Construction of a large-scale image dataset using deep learning with humans in the loop'. In: *arXiv preprint arXiv:1506.03365* (2015).

[109] Mircea Cimpoi et al. ‘Describing textures in the wild’. In: *CVPR*. 2014, pp. 3606–3613.

[110] Catherine Wah et al. ‘The caltech-ucsd birds-200-2011 dataset’. In: (2011).

[111] M-E Nilsback and Andrew Zisserman. ‘A visual vocabulary for flower classification’. In: *CVPR*. Vol. 2. IEEE. 2006, pp. 1447–1454.

[112] Gregory Griffin, Alex Holub and Pietro Perona. ‘Caltech-256 object category dataset’. In: (2007).

[113] Aditya Khosla et al. ‘Novel dataset for fine-grained image categorization: Stanford dogs’. In: *CVPR Workshop on Fine-Grained Visual Categorization (FGVC)*. Vol. 2. 1. Citeseer. 2011.

[114] Dan Hendrycks and Thomas Dietterich. ‘Benchmarking neural network robustness to common corruptions and perturbations’. In: *ICLR*. 2019.

[115] Junkai Huang et al. ‘Trash to Treasure: Harvesting OOD Data with Cross-Modal Matching for Open-Set Semi-Supervised Learning’. In: *ICCV*. 2021, pp. 8310–8319.

[116] Kate Saenko et al. ‘Adapting visual category models to new domains’. In: *ECCV*. Springer. 2010, pp. 213–226.

[117] Xingchao Peng et al. ‘Visda: A synthetic-to-real benchmark for visual domain adaptation’. In: *CVPRW*. 2018, pp. 2021–2026.

[118] Dimitris Tsipras et al. ‘Robustness May Be at Odds with Accuracy’. In: *ICLR*. 2018.

[119] Ian J Goodfellow, Jonathon Shlens and Christian Szegedy. ‘Explaining and harnessing adversarial examples’. In: *ICLR*. 2015.

[120] Feng Liu et al. ‘Probabilistic margins for instance reweighting in adversarial training’. In: *NeurIPS*. Vol. 34. 2021, pp. 23258–23269.

[121] Mingsheng Long et al. ‘Learning transferable features with deep adaptation networks’. In: *ICML*. PMLR. 2015, pp. 97–105.

[122] Xiaobo Xia et al. ‘Are anchor points really indispensable in label-noise learning?’ In: *NeurIPS*. Vol. 32. 2019.

[123] Xiaobo Xia et al. ‘Part-dependent label noise: Towards instance-dependent label noise’. In: *NeurIPS*. Vol. 33. 2020, pp. 7597–7610.

[124] Xiaobo Xia et al. ‘Moderate Coreset: A Universal Method of Data Selection for Real-world Data-efficient Deep Learning’. In: *ICLR*. 2023.

[125] Shiori Sagawa et al. ‘Distributionally robust neural networks for group shifts: On the importance of regularization for worst-case generalization’. In: *International Conference on Learning Representations*. 2020.

[126] Atul Ingle et al. ‘Passive inter-photon imaging’. In: *CVPR*. 2021, pp. 8585–8595.

[127] Chengxi Li et al. ‘Photon-limited object detection using non-local feature matching and knowledge distillation’. In: *CVPR*. 2021, pp. 3976–3987.

[128] Florian Luisier, Thierry Blu and Michael Unser. ‘Image denoising in mixed Poisson–Gaussian noise’. In: *IEEE Transactions on image processing* 20.3 (2010), pp. 696–708.

[129] Klaus E Timmermann and Robert D Nowak. ‘Multiscale modeling and estimation of Poisson processes with application to photon-limited imaging’. In: *IEEE Transactions on Information Theory* 45.3 (1999), pp. 846–862.

[130] Hongseok Namkoong and John C Duchi. ‘Stochastic gradient methods for distributionally robust optimization with f-divergences’. In: *NeurIPS*. Vol. 29. 2016.

[131] Runtian Zhai et al. ‘Doro: Distributional and outlier robust optimization’. In: *ICML*. PMLR. 2021, pp. 12345–12355.

[132] Vihari Piratla, Praneeth Netrapalli and Sunita Sarawagi. ‘Focus on the Common Good: Group Distributional Robustness Follows’. In: *ICLR*. 2021.

[133] Zhenyi Wang et al. ‘Meta-learning without data via wasserstein distributionally-robust model fusion’. In: *UAI*. PMLR. 2022, pp. 2045–2055.

[134] Zhenyi Wang et al. ‘Improving task-free continual learning by distributionally robust memory evolution’. In: *ICML*. PMLR. 2022, pp. 22985–22998.

[135] Nitish Shirish Keskar et al. ‘On large-batch training for deep learning: Generalization gap and sharp minima’. In: *ICLR*. 2017.

[136] Pratik Chaudhari et al. ‘Entropy-sgd: Biasing gradient descent into wide valleys’. In: *ICLR*. 2017.

[137] Pierre Foret et al. ‘Sharpness-aware Minimization for Efficiently Improving Generalization’. In: *ICLR*. 2020.

[138] Dongxian Wu, Shu-Tao Xia and Yisen Wang. 'Adversarial weight perturbation helps robust generalization'. In: *NeurIPS*. Vol. 33. 2020, pp. 2958–2969.

[139] Junbum Cha et al. 'Swad: Domain generalization by seeking flat minima'. In: *Advances in Neural Information Processing Systems*. Vol. 34. 2021, pp. 22405–22418.

[140] Evan Z Liu et al. 'Just train twice: Improving group robustness without training group information'. In: *ICML*. PMLR. 2021, pp. 6781–6792.

[141] Qizhou Wang et al. 'Out-of-distribution Detection with Implicit Outlier Transformation'. In: *ICLR*. 2023.

[142] Shiyu Chang et al. 'Invariant rationalization'. In: *International Conference on Machine Learning*. PMLR. 2020, pp. 1448–1458.

[143] Elliot Creager, Jörn-Henrik Jacobsen and Richard Zemel. 'Environment inference for invariant learning'. In: *ICML*. PMLR. 2021, pp. 2189–2200.

[144] David Krueger et al. 'Out-of-distribution generalization via risk extrapolation (rex)'. In: *ICML*. PMLR. 2021, pp. 5815–5826.

[145] Tatsunori Hashimoto et al. 'Fairness without demographics in repeated loss minimization'. In: *ICML*. PMLR. 2018, pp. 1929–1938.

[146] Klim Kireev, Maksym Andriushchenko and Nicolas Flammarion. 'On the effectiveness of adversarial training against common corruptions'. In: *Uncertainty in Artificial Intelligence*. PMLR. 2022, pp. 1012–1021.

[147] Fabio M Carlucci et al. 'Domain generalization by solving jigsaw puzzles'. In: *CVPR*. 2019, pp. 2229–2238.

[148] Xingchao Peng et al. 'Moment matching for multi-source domain adaptation'. In: *ICCV*. 2019, pp. 1406–1415.

[149] Fengchun Qiao, Long Zhao and Xi Peng. 'Learning to learn single domain generalization'. In: *CVPR*. 2020, pp. 12556–12565.

[150] Yang Shu et al. 'Open domain generalization with domain-augmented meta-learning'. In: *CVPR*. 2021, pp. 9624–9633.

[151] Divyat Mahajan, Shruti Tople and Amit Sharma. 'Domain generalization using causal matching'. In: *ICML*. PMLR. 2021, pp. 7313–7324.

[152] Krikamol Muandet, David Balduzzi and Bernhard Schölkopf. 'Domain generalization via invariant feature representation'. In: *ICML*. PMLR. 2013, pp. 10–18.

[153] Zhuo Huang et al. 'Harnessing Out-Of-Distribution Examples via Augmenting Content and Style'. In: *ICLR*. 2023.

[154] Xingxuan Zhang et al. 'Towards Unsupervised Domain Generalization'. In: *CVPR*. June 2022, pp. 4910–4920.

[155] Ya Li et al. 'Deep domain generalization via conditional invariant adversarial networks'. In: *ECCV*. 2018, pp. 624–639.

[156] Mengyue Yang et al. 'CausalVAE: Disentangled representation learning via neural structural causal models'. In: *CVPR*. 2021, pp. 9593–9602.

[157] Bernhard Schölkopf et al. 'Toward causal representation learning'. In: *Proceedings of the IEEE* 109.5 (2021), pp. 612–634.

[158] Zhongqi Yue et al. 'Transporting causal mechanisms for unsupervised domain adaptation'. In: *ICCV*. 2021, pp. 8599–8608.

[159] Minyoung Kim et al. 'Fisher sam: Information geometry and sharpness aware minimisation'. In: *ICML*. PMLR. 2022, pp. 11148–11161.

[160] Yong Liu et al. 'Towards efficient and scalable sharpness-aware minimization'. In: *CVPR*. 2022, pp. 12360–12370.

[161] Yaowei Zheng, Richong Zhang and Yongyi Mao. 'Regularizing neural networks via adversarial model perturbation'. In: *CVPR*. 2021, pp. 8156–8165.

[162] Aharon Ben-Tal et al. 'Robust solutions of optimization problems affected by uncertain probabilities'. In: *Management Science* 59.2 (2013), pp. 341–357.

[163] John Duchi and Hongseok Namkoong. 'Learning models with uniform performance via distributionally robust optimization'. In: *arXiv preprint arXiv:1810.08750* (2018).

[164] Weihua Hu et al. 'Does distributionally robust supervised learning give robust classifiers?' In: *International Conference on Machine Learning*. PMLR. 2018, pp. 2029–2037.

[165] Dan Hendrycks, Mantas Mazeika and Thomas Dietterich. 'Deep anomaly detection with outlier exposure'. In: *ICLR*. 2019.

[166] Yewen Li et al. ‘Out-of-distribution detection with an adaptive likelihood ratio on informative hierarchical vae’. In: *NeurIPS*. 2022.

[167] Jiawei Du et al. ‘Efficient sharpness-aware minimization for improved training of neural networks’. In: *ICLR*. 2022.

[168] Jiawei Du et al. ‘Sharpness-Aware Training for Free’. In: *arXiv preprint arXiv:2205.14083* (2022).

[169] Zhiyuan Zhang et al. ‘GA-SAM: Gradient-Strength based Adaptive Sharpness-Aware Minimization for Improved Generalization’. In: *arXiv preprint arXiv:2210.06895* (2022).

[170] Yang Zhao, Hao Zhang and Xiuyuan Hu. ‘Penalizing Gradient Norm for Efficiently Improving Generalization in Deep Learning’. In: *ICML*. 2022.

[171] Yang Zhao, Hao Zhang and Xiuyuan Hu. ‘SS-SAM: Stochastic Scheduled Sharpness-Aware Minimization for Efficiently Training Deep Neural Networks’. In: *arXiv preprint arXiv:2203.09962* (2022).

[172] Qihuang Zhong et al. ‘Improving Sharpness-Aware Minimization with Fisher Mask for Better Generalization on Language Models’. In: *EMNLP*. 2022.

[173] Peng Mi et al. ‘Make sharpness-aware minimization stronger: A sparsified perturbation approach’. In: *NeurIPS*. 2022.

[174] Hao Sun et al. ‘AdaSAM: Boosting Sharpness-Aware Minimization with Adaptive Learning Rate and Momentum for Training Deep Neural Networks’. In: *arXiv preprint arXiv:2303.00565* (2023).

[175] Yan Sun et al. ‘Fedspeed: Larger local interval, less communication round, and higher generalization accuracy’. In: *ICLR*. 2023.

[176] Alexandre Rame, Corentin Dancette and Matthieu Cord. ‘Fishr: Invariant gradient variances for out-of-distribution generalization’. In: *ICML*. PMLR. 2022, pp. 18347–18377.

[177] Ishaan Gulrajani and David Lopez-Paz. ‘In search of lost domain generalization’. In: *ICLR*. 2021.

[178] Da Li et al. ‘Deeper, broader and artier domain generalization’. In: *CVPR*. 2017, pp. 5542–5550.

[179] Chaochao Lu et al. ‘Invariant causal representation learning for out-of-distribution generalization’. In: *ICLR*. 2021.

[180] Jonathan Frankle and Michael Carbin. ‘The lottery ticket hypothesis: Finding sparse, trainable neural networks’. In: *ICLR*. 2019.

[181] Jonathan Frankle et al. ‘Linear mode connectivity and the lottery ticket hypothesis’. In: *ICML*. PMLR. 2020, pp. 3259–3269.

[182] Eran Malach et al. ‘Proving the lottery ticket hypothesis: Pruning is all you need’. In: *ICML*. PMLR. 2020, pp. 6682–6691.

[183] Ari Morcos et al. ‘One ticket to win them all: generalizing lottery ticket initializations across datasets and optimizers’. In: *NeurIPS*. Vol. 32. 2019.

[184] Dinghuai Zhang et al. ‘Can subnetwork structure be the key to out-of-distribution generalization?’ In: *ICML*. PMLR. 2021, pp. 12356–12367.

[185] Xiao Zhou et al. ‘Sparse Invariant Risk Minimization’. In: *International Conference on Machine Learning*. PMLR. 2022, pp. 27222–27244.

[186] Biwei Huang et al. ‘Causal discovery from heterogeneous/nonstationary data’. In: *The Journal of Machine Learning Research* 21.1 (2020), pp. 3482–3534.

[187] Raphael Suter et al. ‘Robustly disentangled causal mechanisms: Validating deep representations for interventional robustness’. In: *ICML*. PMLR. 2019, pp. 6056–6065.

[188] Yingbin Bai et al. ‘RSA: Reducing Semantic Shift from Aggressive Augmentations for Self-supervised Learning’. In: *NeurIPS*. Vol. 35. 2022, pp. 21128–21141.

[189] Zhuo Huang et al. ‘Robust Generalization against Photon-Limited Corruptions via Worst-Case Sharpness Minimization’. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2023, pp. 16175–16185.

[190] Qizhou Wang et al. ‘Out-of-distribution Detection with Implicit Outlier Transformation’. In: *ICLR*. 2023.

[191] Hui Kang et al. ‘Unleashing the Potential of Regularization Strategies in Learning with Noisy Labels’. In: *arXiv preprint arXiv:2307.05025* (2023).

[192] Xiaobo Xia et al. ‘Combating Noisy Labels with Sample Selection by Mining High-Discrepancy Examples’. In: *ICCV*. 2023, pp. 1833–1843.

[193] Xiaobo Xia et al. ‘Sample selection with uncertainty of losses for learning with noisy labels’. In: *arXiv preprint arXiv:2106.00445* (2021).

[194] Zhaoqing Wang et al. ‘Exploring set similarity for dense self-supervised representation learning’. In: *CVPR*. 2022, pp. 16590–16599.

[195] Qizhou Wang et al. ‘Learning to Augment Distributions for Out-of-distribution Detection’. In: *NeurIPS*. 2023.

[196] Qizhou Wang et al. ‘Watermarking for Out-of-distribution Detection’. In: *NeurIPS*. 2022.

[197] Chang Liu et al. ‘Learning causal semantic representation for out-of-distribution prediction’. In: *NeurIPS*. Vol. 34. 2021.

[198] Xinwei Sun et al. ‘Recovering Latent Causal Factor for Generalization to Distributional Shifts’. In: *NeurIPS*. Vol. 34. 2021.

[199] Elan Rosenfeld, Pradeep Ravikumar and Andrej Risteski. ‘The risks of invariant risk minimization’. In: *arXiv preprint arXiv:2010.05761* (2020).

[200] Namhoon Lee, Thalaiyasingam Ajanthan and Philip HS Torr. ‘Snip: Single-shot network pruning based on connection sensitivity’. In: *International Conference on Learning Representations*. 2018.

[201] Utku Evci et al. ‘Rigging the lottery: Making all tickets winners’. In: *ICML*. PMLR. 2020, pp. 2943–2952.

[202] Yi-Lin Sung, Varun Nair and Colin A Raffel. ‘Training neural networks with fixed sparse masks’. In: *NeurIPS*. Vol. 34. 2021, pp. 24193–24205.

[203] Tianlong Chen et al. ‘The lottery tickets hypothesis for supervised and self-supervised pre-training in computer vision models’. In: *CVPR*. 2021, pp. 16306–16316.

[204] Tim Dettmers and Luke Zettlemoyer. ‘Sparse networks from scratch: Faster training without losing performance’. In: *arXiv preprint arXiv:1907.04840* (2019).

[205] Róbert Csordás, Sjoerd van Steenkiste and Jürgen Schmidhuber. ‘Are neural nets modular? inspecting functional modularity through differentiable weight masks’. In: *ICLR*. 2020.

[206] Christos Louizos, Max Welling and Diederik P Kingma. ‘Learning Sparse Neural Networks through L_0 Regularization’. In: *ICLR*. 2018.

[207] Shiwei Liu et al. ‘The unreasonable effectiveness of random pruning: Return of the most naive baseline for sparse training’. In: *ICLR*. 2022.

[208] Baochen Sun and Kate Saenko. ‘Deep coral: Correlation alignment for deep domain adaptation’. In: *ECCV*. Springer. 2016, pp. 443–450.

[209] Junbum Cha et al. ‘Domain generalization by mutual-information regularization with pre-trained models’. In: *ECCV*. Springer. 2022, pp. 440–457.

[210] Diederik P Kingma and Jimmy Ba. ‘Adam: A method for stochastic optimization’. In: *arXiv preprint arXiv:1412.6980* (2014).

[211] Jungmin Kwon et al. ‘Asam: Adaptive sharpness-aware minimization for scale-invariant learning of deep neural networks’. In: *ICML*. PMLR. 2021, pp. 5905–5914.

[212] Muhammad Ghifary et al. ‘Domain generalization for object recognition with multi-task autoencoders’. In: *ICCV*. 2015, pp. 2551–2559.

[213] Chen Fang, Ye Xu and Daniel N Rockmore. ‘Unbiased metric learning: On the utilization of multiple datasets and web images for softening bias’. In: *CVPR*. 2013, pp. 1657–1664.

[214] Hemanth Venkateswara et al. ‘Deep hashing network for unsupervised domain adaptation’. In: *CVPR*. 2017, pp. 5018–5027.

[215] Sara Beery, Grant Van Horn and Pietro Perona. ‘Recognition in terra incognita’. In: *ECCV*. 2018, pp. 456–473.

[216] Pang Wei Koh et al. ‘Wilds: A benchmark of in-the-wild distribution shifts’. In: *ICML*. PMLR. 2021, pp. 5637–5664.

[217] Benjamin Recht et al. ‘Do imagenet classifiers generalize to imagenet?’ In: *ICML*. PMLR. 2019, pp. 5389–5400.

[218] Dan Hendrycks et al. ‘The many faces of robustness: A critical analysis of out-of-distribution generalization’. In: *ICCV*. 2021, pp. 8340–8349.

[219] Dan Hendrycks et al. ‘Natural adversarial examples’. In: *CVPR*. 2021, pp. 15262–15271.

[220] Haohan Wang et al. ‘Learning robust global representations by penalizing local predictive power’. In: *NeurIPS*. Vol. 32. 2019.

[221] Andrei Barbu et al. ‘Objectnet: A large-scale bias-controlled dataset for pushing the limits of object recognition models’. In: *NeurIPS*. Vol. 32. 2019.

[222] Alexandre Rame et al. ‘Diverse weight averaging for out-of-distribution generalization’. In: *NeurIPS*. 2022.

[223] Pavel Izmailov et al. ‘Averaging weights leads to wider optima and better generalization’. In: *arXiv preprint arXiv:1803.05407* (2018).

[224] Yingbin Bai and Tongliang Liu. ‘Me-momentum: Extracting hard confident examples from noisily labeled data’. In: *ICCV*. 2021, pp. 9312–9321.

[225] Maksym Andriushchenko and Nicolas Flammarion. ‘Towards understanding sharpness-aware minimization’. In: *ICML*. PMLR. 2022, pp. 639–668.

[226] Alec Radford et al. ‘Learning Transferable Visual Models From Natural Language Supervision’. In: *ICML*. 2021, pp. 8748–8763.

[227] Behrooz Ghorbani, Shankar Krishnan and Ying Xiao. ‘An investigation into neural net optimization via hessian eigenvalue density’. In: *ICML*. PMLR. 2019, pp. 2232–2241.

[228] Alexey Dosovitskiy et al. ‘An image is worth 16x16 words: Transformers for image recognition at scale’. In: *arXiv preprint arXiv:2010.11929* (2020).

[229] Ze Liu et al. ‘Swin transformer: Hierarchical vision transformer using shifted windows’. In: *ICCV*. 2021, pp. 10012–10022.

[230] Wenhai Wang et al. ‘Pyramid vision transformer: A versatile backbone for dense prediction without convolutions’. In: *ICCV*. 2021, pp. 568–578.

[231] Runnan Chen et al. ‘Clip2scene: Towards label-efficient 3d scene understanding by clip’. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2023, pp. 7020–7030.

[232] Yangguang Li et al. ‘Supervision exists everywhere: A data efficient contrastive language-image pre-training paradigm’. In: *arXiv preprint arXiv:2110.05208* (2021).

[233] Junnan Li et al. ‘Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation’. In: *International Conference on Machine Learning*. PMLR. 2022, pp. 12888–12900.

[234] Qizhou Wang et al. ‘Do CLIPs Always Generalize Better than ImageNet Models?’ In: *arXiv preprint arXiv:2403.11497* (2024).

[235] Jiyang Zheng et al. ‘Enhancing Contrastive Learning for Ordinal Regression via Ordinal Content Preserved Data Augmentation’. In: *The Twelfth International Conference on Learning Representations*. 2023.

[236] Yinpeng Dong et al. ‘Benchmarking robustness of 3d object detection to common corruptions’. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2023, pp. 1022–1032.

[237] Lingdong Kong et al. ‘Robo3d: Towards robust and reliable 3d perception against corruptions’. In: *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 2023, pp. 19994–20006.

[238] Zhuo Huang et al. ‘Winning prize comes from losing tickets: Improve invariant learning by exploring variant parameters for out-of-distribution generalization’. In: *International Journal of Computer Vision* (2024), pp. 1–19.

[239] Ziming Hong et al. ‘Improving Non-Transferable Representation Learning by Harnessing Content and Style’. In: *The Twelfth International Conference on Learning Representations*. 2024. URL: https://openreview.net/forum?id=FYKVPOHCpE.

[240] Xidong Peng et al. ‘SAM-guided Unsupervised Domain Adaptation for 3D Segmentation’. In: *arXiv preprint arXiv:2310.08820* (2023).

[241] Zijian Zhu et al. ‘Understanding the Robustness of 3D Object Detection With Bird’s-Eye-View Representations in Autonomous Driving’. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2023, pp. 21600–21610.

[242] Yang Shu et al. ‘CLIPood: Generalizing CLIP to Out-of-Distributions’. In: *ICML*. PMLR. 2023.

[243] Sachin Goyal et al. ‘Finetune like you pretrain: Improved finetuning of zero-shot vision models’. In: *CVPR*. 2023, pp. 19338–19347.

[244] Mitchell Wortsman et al. ‘Robust fine-tuning of zero-shot models’. In: *CVPR*. 2022, pp. 7959–7971.

[245] Zhuo Huang et al. ‘Towards out-of-modal generalization without instance-level modal correspondence’. In: *The Thirteenth International Conference on Learning Representations*. 2025.

[246] Jean-Baptiste Alayrac et al. ‘Flamingo: a visual language model for few-shot learning’. In: *NeurIPS*. Vol. 35. 2022, pp. 23716–23736.

[247] Anas Awadalla et al. ‘Openflamingo: An open-source framework for training large autoregressive vision-language models’. In: *arXiv preprint arXiv:2308.01390* (2023).

[248] Runnan Chen et al. ‘Towards label-free scene understanding by vision foundation models’. In: *Advances in Neural Information Processing Systems* 36 (2024).

[249] Tao Gong et al. ‘Multimodal-gpt: A vision and language model for dialogue with humans’. In: *arXiv preprint arXiv:2305.04790* (2023).

[250] Junnan Li et al. ‘Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models’. In: *arXiv preprint arXiv:2301.12597* (2023).

[251] Bo Li et al. ‘Otter: A multi-modal model with in-context instruction tuning’. In: *arXiv preprint arXiv:2305.03726* (2023).

[252] Haotian Liu et al. ‘Visual Instruction Tuning’. In: *NeurIPS*. 2023.

[253] Qinghao Ye et al. ‘mplug-owl: Modularization empowers large language models with multimodality’. In: *arXiv preprint arXiv:2304.14178* (2023).

[254] Deyao Zhu et al. ‘Minigpt-4: Enhancing vision-language understanding with advanced large language models’. In: *arXiv preprint arXiv:2304.10592* (2023).

[255] Tom Brown et al. ‘Language models are few-shot learners’. In: *Advances in neural information processing systems* 33 (2020), pp. 1877–1901.

[256] Hyung Won Chung et al. ‘Scaling instruction-finetuned language models’. In: *arXiv preprint arXiv:2210.11416* (2022).

[257] Luciano Floridi and Massimo Chiriatti. ‘GPT-3: Its nature, scope, limits, and consequences’. In: *Minds and Machines* 30 (2020), pp. 681–694.

[258] OpenAI. *GPT-4 Technical Report*. 2023. arXiv: 2303.08774 [cs.CL].

[259] Teven Le Scao et al. ‘Bloom: A 176b-parameter open-access multilingual language model’. In: *arXiv preprint arXiv:2211.05100* (2022).

[260] Hugo Touvron et al. ‘Llama: Open and efficient foundation language models’. In: *arXiv preprint arXiv:2302.13971* (2023).

[261] Hugo Touvron et al. ‘Llama 2: Open foundation and fine-tuned chat models’. In: *arXiv preprint arXiv:2307.09288* (2023).

[262] Lianmin Zheng et al. *Judging LLM-as-a-judge with MT-Bench and Chatbot Arena*. 2023. arXiv: 2306.05685 [cs.CL].

[263] Puning Yang et al. ‘Exploring Criteria of Loss Reweighting to Enhance LLM Unlearning’. In: *Forty-second International Conference on Machine Learning*. 2025. URL: https://openreview.net/forum?id=mGOugCZlAq.

[264] Jianfeng Wang et al. ‘Git: A generative image-to-text transformer for vision and language’. In: *arXiv preprint arXiv:2205.14100* (2022).

[265] Shaohan Huang et al. ‘Language is not all you need: Aligning perception with language models’. In: *arXiv preprint arXiv:2302.14045* (2023).

[266] Yuexiang Zhai et al. ‘Investigating the catastrophic forgetting in multimodal large language models’. In: *arXiv preprint arXiv:2309.10313* (2023).

[267] Bo Han et al. ‘Co-teaching: Robust training of deep neural networks with extremely noisy labels’. In: *NeurIPS*. Vol. 31. 2018.

[268] Tongliang Liu and Dacheng Tao. ‘Classification with noisy labels by importance reweighting’. In: *IEEE Transactions on pattern analysis and machine intelligence* 38.3 (2015), pp. 447–461.

[269] Yexiong Lin et al. ‘Do We Need to Penalize Variance of Losses for Learning with Label Noise?’ In: *arXiv preprint arXiv:2201.12739* (2022).

[270] Yexiong Lin et al. ‘CS-Isolate: Extracting Hard Confident Examples by Content and Style Isolation’. In: *Thirty-seventh Conference on Neural Information Processing Systems*. 2023.

[271] Nagarajan Natarajan et al. ‘Learning with noisy labels’. In: *NeurIPS*. Vol. 26. 2013.

[272] Yuhao Wu et al. ‘Mitigating Label Noise on Graph via Topological Sample Selection’. In: *arXiv preprint arXiv:2403.01942* (2024).

[273] Yuhao Wu et al. ‘Making Binary Classification from Multiple Unlabeled Datasets Almost Free of Supervision’. In: *arXiv preprint arXiv:2306.07036* (2023).

[274] Xiaobo Xia et al. ‘Robust early-learning: Hindering the memorization of noisy labels’. In: *International conference on learning representations*. 2020.

[275] Yu Yao et al. ‘Dual t: Reducing estimation error for transition matrix in label-noise learning’. In: *Advances in neural information processing systems* 33 (2020), pp. 7260–7271.

[276] Yu Yao et al. ‘Which is better for learning with noisy labels: the semi-supervised method or modeling label noise?’ In: *International Conference on Machine Learning*. PMLR. 2023, pp. 39660–39673.

[277] Suqin Yuan, Lei Feng and Tongliang Liu. ‘Early Stopping Against Label Noise Without Validation Data’. In: *The Twelfth International Conference on Learning Representations*. 2024. URL: https://openreview.net/forum?id=CMzF2aOfqp.

[278] Yixin Chen et al. ‘Lightweight In-Context Tuning for Multimodal Unified Models’. In: *arXiv preprint arXiv:2310.05109* (2023).

[279] Bo Li et al. ‘MIMIC-IT: Multi-Modal In-Context Instruction Tuning’. In: (2023). arXiv: 2306.05425 [cs.CV].

[280] Michihiro Yasunaga et al. ‘Retrieval-augmented multimodal language modeling’. In: *ICML*. 2023.

[281] Haozhe Zhao et al. ‘MMICL: Empowering Vision-language Model with Multi-Modal In-Context Learning’. In: *arXiv preprint arXiv:2309.07915* (2023).

[282] Jia Deng et al. ‘Imagenet: A large-scale hierarchical image database’. In: *2009 IEEE conference on computer vision and pattern recognition*. Ieee. 2009, pp. 248–255.

[283] Zhuo Huang et al. ‘Winning prize comes from losing tickets: Improve invariant learning by exploring variant parameters for out-of-distribution generalization’. In: *arXiv preprint arXiv:2310.16391* (2023).

[284] Dawei Zhou et al. ‘Towards defending against adversarial examples via attack-invariant features’. In: *International conference on machine learning*. PMLR. 2021, pp. 12835–12845.

[285] Dawei Zhou et al. ‘Improving adversarial robustness via mutual information estimation’. In: *International Conference on Machine Learning*. PMLR. 2022, pp. 27338–27352.

[286] Terrance DeVries and Graham W Taylor. ‘Improved regularization of convolutional neural networks with cutout’. In: *arXiv preprint arXiv:1708.04552* (2017).

[287] Sangdoo Yun et al. ‘Cutmix: Regularization strategy to train strong classifiers with localizable features’. In: *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 2019, pp. 6023–6032.

[288] Hongyi Zhang et al. ‘mixup: Beyond empirical risk minimization’. In: *ICLR*. 2017.

[289] Dan Hendrycks, Mantas Mazeika and Thomas Dietterich. ‘Deep anomaly detection with outlier exposure’. In: *arXiv preprint arXiv:1812.04606* (2018).

[290] Wenjin Hou et al. ‘Visual-Augmented Dynamic Semantic Prototype for Generative Zero-Shot Learning’. In: *arXiv preprint arXiv:2404.14808* (2024).

[291] Wei-Lin Chiang et al. *Vicuna: An Open-Source Chatbot Impressing GPT-4 with 90%* ChatGPT Quality*. Mar. 2023. URL: https://lmsys.org/blog/2023-03-30-vicuna/.

[292] Susan Zhang et al. ‘Opt: Open pre-trained transformer language models’. In: *arXiv preprint arXiv:2205.01068* (2022).

[293] Wenliang Dai et al. ‘InstructBLIP: Towards General-purpose Vision-Language Models with Instruction Tuning’. In: *arXiv preprint arXiv:2305.06500* (2023).

[294] Edward J Hu et al. ‘Lora: Low-rank adaptation of large language models’. In: *arXiv preprint arXiv:2106.09685* (2021).

[295] Rohit Girdhar et al. ‘ImageBind: One Embedding Space To Bind Them All’. In: *CVPR*. 2023.

[296] Yixuan Su et al. ‘PandaGPT: One Model To Instruction-Follow Them All’. In: *arXiv preprint arXiv:2305.16355* (2023).

[297] Muyang Li et al. ‘InstanT: Semi-supervised Learning with Instance-dependent Thresholds’. In: *Advances in Neural Information Processing Systems*. Vol. 36. 2024.

[298] Sang Michael Xie et al. ‘An explanation of in-context learning as implicit bayesian inference’. In: *arXiv preprint arXiv:2111.02080* (2021).

[299] Yuxin Fang et al. ‘EVA: Exploring the Limits of Masked Visual Representation Learning at Scale’. In: *arXiv preprint arXiv:2211.07636* (2022).

[300] Yann LeCun et al. ‘Gradient-based learning applied to document recognition’. In: *Proceedings of the IEEE* 86.11 (1998), pp. 2278–2324.

[301] Yinpeng Dong et al. ‘Viewfool: Evaluating the robustness of visual recognition to adversarial viewpoints’. In: *Advances in Neural Information Processing Systems* 35 (2022), pp. 36789–36803.

[302] Pang Wei Koh et al. ‘WILDS: A Benchmark of in-the-Wild Distribution Shifts’. In: *International Conference on Machine Learning (ICML)*. 2021.

[303] Ishaan Gulrajani and David Lopez-Paz. ‘In search of lost domain generalization’. In: *arXiv preprint arXiv:2007.01434* (2020).

[304] Xiaohua Zhai et al. ‘Scaling vision transformers’. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2022, pp. 12104–12113.

[305] Aengus Lynch et al. ‘Spawrious: A benchmark for fine control of spurious correlation biases’. In: *arXiv preprint arXiv:2303.05470* (2023).

[306] Ziwei Liu et al. ‘Deep Learning Face Attributes in the Wild’. In: *Proceedings of International Conference on Computer Vision (ICCV)*. Dec. 2015.

[307] Qizhou Wang et al. ‘Learning to augment distributions for out-of-distribution detection’. In: *Advances in Neural Information Processing Systems*. Vol. 36. 2024.

[308] Kaiming He et al. ‘Mask r-cnn’. In: *Proceedings of the IEEE international conference on computer vision*. 2017, pp. 2961–2969.

[309] Shaoqing Ren et al. ‘Faster r-cnn: Towards real-time object detection with region proposal networks’. In: *NeurIPS* 28 (2015).

[310] Jacob Devlin et al. ‘Bert: Pre-training of deep bidirectional transformers for language understanding’. In: *arXiv preprint arXiv:1810.04805* (2018).

[311] Zhuo Huang et al. ‘Machine Vision Therapy: Multimodal Large Language Models Can Enhance Visual Robustness via Denoising In-Context Learning’. In: *Forty-first International Conference on Machine Learning*. 2024. URL: https://openreview.net/forum?id=LwOfVWgEzS.

[312] Jiquan Ngiam et al. ‘Multimodal deep learning’. In: *ICML*. 2011, pp. 689–696.

[313] Richard Socher et al. ‘Zero-shot learning through cross-modal transfer’. In: *NeurIPS* 26 (2013).

[314] Haoyu Wang et al. ‘Noisegpt: Label noise detection and rectification through probability curvature’. In: *NeurIPS*. 2024.

[315] Xiaohao Liu et al. ‘Towards modality generalization: A benchmark and prospective analysis’. In: *Proceedings of the 33rd ACM International Conference on Multimedia*. 2025, pp. 12179–12188.

[316] Yu Huang et al. ‘What makes multi-modal learning better than single (provably)’. In: *NeurIPS*. Vol. 34. 2021, pp. 10944–10956.

[317] Zhou Lu. ‘A theory of multimodal learning’. In: *NeurIPS* 36 (2024).

[318] Xinwei Sun et al. ‘Tcgm: An information-theoretic framework for semi-supervised multi-modality learning’. In: *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part III 16*. Springer. 2020, pp. 171–188.

[319] Han Zhang et al. ‘Cross-modal contrastive learning for text-to-image generation’. In: *CVPR*. 2021, pp. 833–842.

[320] Liangli Zhen et al. ‘Deep supervised cross-modal retrieval’. In: *CVPR*. 2019, pp. 10394–10403.

[321] Maja Pantic and Leon JM Rothkrantz. ‘Toward an affect-sensitive multimodal human-computer interaction’. In: *Proceedings of the IEEE* 91.9 (2003), pp. 1370–1390.

[322] Muhammad Arifur Rahman et al. ‘Explainable multimodal machine learning for engagement analysis by continuous performance test’. In: *International Conference on Human-Computer Interaction*. Springer. 2022, pp. 386–399.

[323] Yunfan Jiang et al. ‘VIMA: Robot Manipulation with Multimodal Prompts’. In: *ICML*. PMLR. 2023, pp. 14975–15022.

[324] Youngjae Yu et al. ‘Fusing pre-trained language models with multimodal prompts through reinforcement learning’. In: *CVPR*. 2023, pp. 10845–10856.

[325] Masukichi Hashimoto. ‘Origin of the Compass’. In: *Memoirs of the Research Department of the Toyo Bunko (The Oriental Library)* 1 (1926), pp. 69–92.

[326] Heidi E Harley, Erika A Putman and Herbert L Roitblat. 'Bottlenose dolphins perceive object features through echolocation'. In: *Nature* 424.6949 (2003), pp. 667–669.

[327] LA Weinstein. 'Electromagnetic waves'. In: *Radio i svyaz', Moscow* (1988).

[328] Manolis Savva et al. 'Habitat: A platform for embodied ai research'. In: *ICCV*. 2019, pp. 9339–9347.

[329] Zehan Wang et al. 'FreeBind: Free Lunch in Unified Multimodal Space via Knowledge Fusion'. In: *ICML*. 2024.

[330] Bin Zhu et al. 'Languagebind: Extending video-language pretraining to n-modality by language-based semantic alignment'. In: *arXiv preprint arXiv:2310.01852* (2023).

[331] Paul Pu Liang et al. 'Multimodal learning without labeled multimodal data: Guarantees and applications'. In: *arXiv preprint arXiv:2306.04539* (2023).

[332] Paul Pu Liang et al. 'Cross-modal generalization: Learning in low resource modalities via meta-alignment'. In: *Proceedings of the 29th ACM International Conference on Multimedia*. 2021, pp. 2680–2689.

[333] Yan Xia et al. 'Achieving Cross Modal Generalization with Multimodal Unified Representation'. In: *NeurIPS* 36 (2024).

[334] Xiaohao Liu et al. 'Towards Modality Generalization: A Benchmark and Prospective Analysis'. In: *arXiv preprint arXiv:2412.18277* (2024).

[335] Junhong Shen et al. 'Cross-modal fine-tuning: Align then refine'. In: *ICML*. PMLR. 2023, pp. 31030–31056.

[336] Lincan Cai et al. 'Enhancing Cross-Modal Fine-Tuning with Gradually Intermediate Modality Generation'. In: *arXiv preprint arXiv:2406.09003* (2024).

[337] Shuang Ma, Daniel McDuff and Yale Song. 'Unpaired image-to-speech synthesis with multimodal information bottleneck'. In: *ICCV*. 2019, pp. 7598–7607.

[338] Ying Wang, Tim GJ Rudner and Andrew G Wilson. 'Visual explanations of image-text representations via multi-modal information bottleneck attribution'. In: *NeurIPS* 36 (2023), pp. 16009–16027.

[339] Yingying Fang et al. 'Dynamic Multimodal Information Bottleneck for Multimodality Classification'. In: *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*. 2024, pp. 7696–7706.

[340] Hao Dong et al. 'SimMMDG: A simple and effective framework for multi-modal domain generalization'. In: *NeurIPS* 36 (2023), pp. 78674–78695.

[341] Zihui Xue et al. 'The modality focusing hypothesis: Towards understanding cross-modal knowledge distillation'. In: *arXiv preprint arXiv:2206.06487* (2022).

[342] Paul L Williams and Randall D Beer. 'Nonnegative decomposition of multivariate information'. In: *arXiv preprint arXiv:1004.2515* (2010).

[343] Alexander A Alemi et al. 'Deep variational information bottleneck'. In: *arXiv preprint arXiv:1612.00410* (2016).

[344] Karthik Sridharan and Sham M Kakade. 'An information theoretic framework for multi-view learning'. In: *COLT*. 114. 2008, pp. 403–414.

[345] Letian Fu et al. 'A Touch, Vision, and Language Dataset for Multimodal Alignment'. In: *ICML*. 2024. URL: https://openreview.net/forum?id=tFEOOH9eH0.

[346] Xinyu Jia et al. 'LLVIP: A visible-infrared paired dataset for low-light vision'. In: *ICCV*. 2021, pp. 3496–3504.

[347] Nathan Silberman et al. 'Indoor segmentation and support inference from rgbd images'. In: *Computer Vision–ECCV 2012: 12th European Conference on Computer Vision, Florence, Italy, October 7-13, 2012, Proceedings, Part V 12*. Springer. 2012, pp. 746–760.

[348] Honglie Chen et al. 'Vggsound: A large-scale audio-visual dataset'. In: *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE. 2020, pp. 721–725.

[349] Jun Xu et al. 'Msr-vtt: A large video description dataset for bridging video and language'. In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2016, pp. 5288–5296.

[350] AmirAli Bagher Zadeh et al. 'Multimodal language analysis in the wild: Cmu-mosei dataset and interpretable dynamic fusion graph'. In: *Proceedings of the 56th Annual*

Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). 2018, pp. 2236–2246.

[351] Yves Grandvalet and Yoshua Bengio. 'Semi-supervised learning by entropy minimization'. In: *NeurIPS* 17 (2004).

[352] Kaiming He et al. 'Momentum contrast for unsupervised visual representation learning'. In: *CVPR*. 2020, pp. 9729–9738.

[353] Junchi Yang et al. 'Faster single-loop algorithms for minimax optimization without strong concavity'. In: *AISTATS*. PMLR. 2022, pp. 5485–5517.

[354] Maher Nouiehed et al. 'Solving a class of non-convex min-max games using iterative first order methods'. In: *NeurIPS*. Vol. 32. 2019.