Removing Bias for Out-of-Distribution Generalization

Jiaxin Qi

School of Computer Science and Engineering

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

2023

Statement of Originality

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

15/12/2022

.....

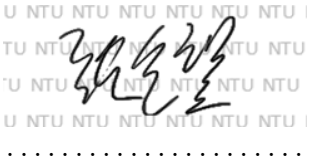
Date

.....

Jiaxin Qi

Supervisor Declaration Statement

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



15/12/2022

.....

Date

.....

Asst. Prof. Hanwang Zhang

Authorship Attribution Statement

This thesis contains materials from 2 papers published in the following peer-reviewed conferences and 2 papers that are currently submitted to a conference in which I am listed as the first author.

Chapter 3 is published as **Jiaxin Qi**, Yulei Niu, Jianqiang Huang, and Hanwang Zhang. Two Causal Principles for Improving Visual Dialog. *CVPR 2020*.

The contributions of the co-authors are as follows:

- Prof. Hanwang Zhang leads the research direction and participates in the paper polishing.
- Yulei Niu participates in the idea discussion and paper polishing.
- I design the algorithm of principles, take charge of the whole code implementation, and write the draft paper.

Chapter 4 is submitted to CVPR 2023: **Jiaxin Qi***, Zike Wu*, Qianru Sun, Xian-Sheng Hua, and Hanwang Zhang. DOMAIN+: Splitting a New Influential Domain for Domain Generalization.[1] *Under Review*.

The contributions of the co-authors are as follows:

- Prof. Hanwang Zhang designs the research topic and joins the paper writing.
- Prof. Qianru Sun polishes the paper.
- Co-first author Zike Wu participates in the idea discussion and polishes the paper, and takes charge of the code implementation.
- I participate in the idea discussion and algorithm design, take charge of part of the code implementation, and write the draft paper.

Chapter 5 contains two parts, the first one is submitted to CVPR 2023: **Jiaxin Qi**, Qianru Sun, Xian-Sheng Hua, and Hanwang Zhang. A Two-stage Method for Training Unbiased Models. *Under Review*.

The contributions of the co-authors are as follows:

- Prof. Hanwang Zhang guides the research direction and participates in the paper writing and polishing.
- Prof. Qianru Sun polishes the paper.
- I design the algorithm, take charge of the whole code implementation and write the draft paper.

[1]The superscript $*$ indicates the equal contributions

The second part of Chapter 5 is published as **Jiaxin Qi**, Kaihua Tang, Qianru Sun, Xian-Sheng Hua, and Hanwang Zhang. Class Is Invariant to Context and Vice Versa: On Learning Invariance for Out-Of-Distribution Generalization. *ECCV 2022*.

The contributions of the co-authors are as follows:

- Prof. Hanwang Zhang guides the research direction and participates in the paper writing and polishing.
- Kaihua Tang participates in the discussion of the algorithm and Prof. Qianru Sun participates in the paper polishing.
- I design the algorithm, take charge of the full code implementation and write the paper.

15/12/2022

.....

Date

.....

Jiaxin Qi

Acknowledgements

First of all, I would like to express my deep gratitude to my supervisor, Prof. Hanwang Zhang. Without his patient guidance and constant help, I would not have been able to complete my research and studies during my Ph.D. Whenever I think back to the past four years, I can always find that he has been helping and witnessing my path. When I felt lost at the end of my undergraduate studies, he gave me the opportunity to pursue a Ph.D. and supported my enrollment; when I encountered difficulties in my research and life, he gave me encouragement and necessary help; when I almost gave up on some projects, he gave me the strength to brighten my future. He is a great mentor in research and life, his ideas are always professional and correct, and he always waits patiently for me to keep up with him, but unfortunately, my ability is limited and I always fail to meet his expectations, for which I always feel sorry to him. He supports every decision we make and respects our choices. I believe that in the future, even if I don't do the relevant research, I can always think of his words to me.

I would like to acknowledge my collaborators, Yulei Niu, Kaihua Tang, Zike Wu, Jianqiang Huang, Prof. Qianru Sun, etc., for their support and insightful comments, I would like to thank my friends I met in MReal Lab, they are Xu Yang, Daqing Liu, Dong Zhang, Xinting Hu, Zhongqi Yue, Tan Wang, Beier Zhu, Yucheng Han, Xuanyu Yi, etc. I also would like to thank my girlfriend, Xiaohan Yi for her support. I would like to thank all my mentors and friends during my internship at Alibaba DAMO Academy: Zhou Fang, Yao Liu, Yu Qin, and Xian-Sheng Hua.

Finally, I would like to express my deepest gratitude to my parents Feng Qi and Hong Wang, and my family. I am grateful that they have been supportive of my studies and life and willing to let me make my own decisions.

Jiaxin Qi, December 2022

Abstract

Deep models have a strong ability to fit the training data, and thus can achieve high performance when the testing data is sampled from the same distribution as the training. However, in practice, the deep models fail to perform perfectly because the testing data is usually Out-of-Distribution (OOD) compared to the training, which is known as the OOD Generalization problem. The underlying reason is that, in the training, besides the causal effect, *i.e.*, the causalities between inputs and outputs which describe the data generation process and will not change under any data distribution, the models also learn the bias, *i.e.*, the spurious correlations between inputs and outputs which only exists in the training distribution, and thus learning such bias will make the model fail to generalize to OOD data.

To help the models achieve better OOD Generalization performance, we need to pursue the causal effect by removing the learned bias. However, due to the various data organization formats and different given inputs, it is hard to propose a uniform bias removal strategy, and thus we categorize the OOD Generalization tasks into three camps and conduct specific case studies for each one: 1) OOD Generalization with Multiple Modalities, where multiple modalities, such as language and image, are provided in the training, and we focus on a specific case, *Visual Dialog*, to analyze its underlying causal relationships between the modalities and propose two causal principles to remove the history bias and user bias for better OOD performance. 2) OOD Generalization with Multiple Domains, where there is only one modality, images, but multiple training domains and their domain annotations are given. We focus on *Domain Generalization* (DG) and propose to create a new domain by cross-domain influence to remove the "spurious invariance" bias to help current DG methods achieve better OOD performance. 3) OOD Generalization with no Additional Annotations, where only one modality, images, and one training domain with no additional annotations, such as domain annotations or bias annotations, are given in the training. We focus on a specific case, *Debiasing*, and propose two algorithms for removing bias. First, we design a two-stage

pipeline with re-weighting methods to effectively remove the underlying context bias. Second, due to the context estimation method used by current re-weighting is hard to succeed when class effect and context effect are entangled, we propose Invariant Risk Minimization for Context to disentangle the context to achieve better re-weighting for removing context bias to achieve better OOD Generalization for debiasing.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Motivations

Given a training set of inputs and outputs, *e.g.*, images and corresponding class labels, the models are trained to infer the corresponding outputs when given some inputs, *e.g.*, learning to predict class labels for input images. The classical approach to optimize the models is to minimize some predefined losses, *e.g.*, minimizing the cross-entropy loss for classification, while such optimization allows anything, that can reduce the training loss, to be learned by the models through their powerful learning ability [13]. For example, in the classification problem for cats and dogs, if most cats are white and most dogs are black in the training set, the models will inevitably learn white and black to help them classify the cats and dogs, respectively, due to such learning can decrease the training loss and thus be encouraged. However, a model that uses white to classify cats will fail when there are more black cats in the testing, which is the **Out-of-Distribution (OOD) Generalization** problem [11, 12, 14–18], indicating the testing distribution, *i.e.*, more black cats, is "out-of" the training distribution, *i.e.*, more white cats. In practice, any shifts of class-irrelevant attributes between training and testing, such as texture, position, background, or other hardly noticeable attributes can induce the OOD Generalization problem.

Intuitively, we say that it is incorrect to classify cats and dogs by using color, which comes from our thinking about the nature of the data generation process, *i.e.*, a cat is labeled as "cat" not because it is white, but because it has the structure

and morphology of cat, and we call them class features (or causal features because causality indeed describes the nature of data generation [19] and see Section 3.2 for more details), which is the golden rule for classifying cats and dogs in any data distribution. In contrast, if other features have correlations with class labels in the training, they are called spurious correlations because such correlations only exist in the training distribution, and the underlying reason for OOD Generalization is that the models learn such spurious correlations for classification, which is called bias, *e.g.*, using white to classify cats due to there are more white cats in the training. Therefore, to achieve better OOD Generalization performance, we need to remove the bias learned by the models.

Due to the various data organization formats and different given inputs of OOD Generalization tasks, it is hard to propose a uniform algorithm for removing bias. Therefore, we categorize them into three camps and conduct specific case studies for each one. We first introduce the characteristics and our ideas for removing bias for each camp: **1) OOD Generalization with Multiple Modalities.** Because there are multiple modalities in the training, it is beneficial for us to analyze the underlying causal relationships between them to find the bias in each modality or across some modalities. Therefore, we focus on Visual Dialog and analyze its causal graph to find the overlooked bias of the traditional methods and remove it. **2) OOD Generalization with Multiple Domains.** Because multiple domains and domain annotations are given in the training, case, we can utilize the traditional Domain Generalization methods, where domain annotations are required, to remove the domain-specific bias, and we focus on improving them by further removing the "spurious invariance" bias. **3) OOD Generalization with no Additional Annotations.** This is the most challenging camp because no additional annotations are given to help us remove the bias, and thus we need to find the bias in the training process. Because the re-weighting methods are usually implemented to remove the context bias in the case, Debiasing, we propose two algorithms to realize better re-weighting. In the following, we will briefly introduce our motivations and contributions to each camp.

OOD Generalization with Multiple Modalities. For this camp, we focus on Visual Dialog and analyze it by causal graph paradigm and find the underlying bias which degenerates the existing methods. For removing the bias, we present two model-agnostic causal principles, according to our careful analysis of the causality

behind the VisDial data generation process and model learning process, which are overlooked by traditional methods. Specifically, Principle 1 says that the direct input of the dialog history should be removed for the decoder, *i.e.*, answering model, otherwise the shortcut will introduce the harmful history bias. Principle 2 suggests that an unobserved confounder for history, question, and answer induces spurious correlations from the training data, and should be added to the traditional VisDial causal graph, which introduces user bias to the models and should be de-confounded. In the implementation, for Principle 1, we can easily cut the direct feeding of history into the answer generation process of the decoder; for Principle 2, we implement de-confounding methods with our designed losses to remove the confounding bias. Our implementations for the proposed causal principles are model-agnostic, and thus are applicable to any traditional method, and we also conduct extensive experiments to show the effectiveness of removing the bias to improve the OOD performance of our implementations.

OOD Generalization with Multiple Domains. For this camp, we focus on Domain Generalization (DG), where we first analyze the bias in DG and point out that all current methods suffer the bias caused by "spurious invariance", which is invariant between the training domains but not in the testing domain. According to our careful analysis, the reason for learning such bias is that the contribution of the minority training samples without such spurious invariance is outgunned. Therefore, we present a cross-domain influence-based method, DOMAIN+, to split these samples out of the original domains to form a new domain, and then the spurious invariance is no longer invariant and removed. In practice, motivated by Influence Function, we estimate sample influence by up-weighting the sample and then calculating how much the invariance loss of the other training domains changes—the more it changes, the higher the influence of the sample and the more likely it should be split out. Then, with the new domain, we can improve any DG methods to achieve better OOD performance, and we conduct extensive experiments to justify our statements.

OOD Generalization with no Additional Annotations. For this camp, we focus on Debiasing, where we first analyze the context bias and find two weaknesses of the existing methods, the "anti-training bias" bias in the current one-stage reweighting methods and the improper estimation for context, inducing inaccurate sample-wise weights for reweighting. Therefore, in the first part, we present a

simple yet effective two-stage method, dubbed Two, containing biased training and unbiased finetuning, which is fundamentally different from existing debiasing methods, which are specially designed to counter the training bias but fall into another "anti-training bias" bias in testing, and thus they are essentially still biased. We introduce the pipeline of our Two and provide detailed justifications for its design philosophy, and we also conduct various experiments to demonstrate the effectiveness of our proposed algorithms. In the second part, we first argue that the widely adopted assumption in prior work, the context bias can be directly annotated or estimated from biased class prediction, rendering the context incomplete or even incorrect. In contrast, we point out the ever-overlooked principle, context is invariant to class, which motivates us to consider the classes as the varying environments as regularization to extract better context features. We implement this idea by minimizing the contrastive loss of intra-class sample similarity while assuring this similarity is invariant across all classes, namely IRMCon. On benchmarks with various debiasing datasets, we show that a simple re-weighting-based classifier equipped with our context estimator achieves state-of-the-art OOD performance.

1.2 Outline of the Thesis

In Chapter 1, we briefly introduce the reasons for the OOD Generalization problem, *i.e.*, the bias, and then we briefly introduce the different OOD Generalization camps and our motivations for each camp.

In Chapter 2, we systematically review the OOD Generalization problem, categorize them into three camps and introduce some specific cases, and then we provide details for them. Then, we introduce the related technologies that inspire our algorithms and implementations.

In Chapter 3, we focus on Visual Dialog (VisDial) for OOD Generalization with Multiple Modalities. We first introduce the causal theory and analyze VisDial by causal graph paradigm to find the underlying bias which degenerates the existing methods. Then, we propose two model-agnostic causal principles to help the current method remove bias and achieve better OOD performance, which is demonstrated by our experiments.

In Chapter 4, we focus on Domain Generalization (DG) for OOD Generalization with Multiple Domains. we first analyze the bias in DG and point out that all current methods suffer the "spurious invariance" bias, which is caused by rare samples. Then, we propose a cross-domain influence-based method, DOMAIN+, to split rare samples to form a new domain, which is model-agnostic and can improve current DG methods to achieve better OOD performance, which is also demonstrated by our experiments.

In Chapter 5, we focus on Debiasing for OOD Generalization with no Additional Annotations. We propose two algorithms to remove the context bias in Debiasing. In the first part, we point out current re-weighting methods suffer the "anti-training bias" bias in testing and present a simple yet effective two-stage method, dubbed TWO, which achieves SOTA performance under different training/testing distribution combinations. In the second part, we point out that the current re-weighting methods improperly estimate the context by a classification objective and propose our IRMCon by minimizing the contrastive loss of intra-class sample similarity while assuring this similarity is invariant across all classes, and we also conduct extensive experiments to demonstrate the effectiveness of our algorithms.

In Chapter 6, we summarize our contributions to remove bias for different OOD Generalization cases and finally propose our future work.

Chapter 2

Literature Review

2.1 Out-of-Distribution (OOD) Generalization

The assumption of independent and identical distribution (IID) for training and testing is fundamental to the success of traditional deep models, that under the assumption, the model generalization emerges easily [20]. However, in practice, the data distribution is always shifted, which is known as the Out-of-Distribution (OOD) Generalization problem, inducing a dramatic drop in the performance of deep models [2, 21–23]. In general, any test distribution that has not been seen in training can be interpreted as an OOD Generalization task, and we can categorize them into different camps according to the given conditions: **1) OOD Generalization with multiple modalities**, such as *Visual Dialog* (VisDial) [24–26], where multiple modalities, *i.e.*, an image, a dialog history, and a current question, are given for models in the training to learn to ask the questions and form a dialog with humans for the given image. It is OOD due to in the training only one answer is given to the corresponding question to train the model, but in the testing, the model needs to produce many suitable answers; *Visual Question Answering* (VQA) [27–29], where an image and the corresponding question is given for model to learn to answer the question, which can be seen as one-round Visual Dialog. **2) OOD Generalization with multiple domains**, such as Domain Generalization [11, 18, 30–32], where the model is trained on multiple different domains with domain annotations to generalize to any unseen domain. **3) OOD Generalization with no additional annotations**, which is the most common one,

such as *Debiasing* [14, 33–36], where a class-balanced training set is given and the model needs to debias any underlying bias existed in the training and generalize to an unbiased testing set. In this thesis, we mainly focus on *Debiasing* as our case study for OOD Generalization with no additional annotations. However, there are several other popular cases in this camp with some different settings compared to *Debiasing*. We briefly introduce these cases and provide the comparisons: *Domain Adaptation* [15, 37–40], where only one training domain and test domain images without class labels are given in the training and the model is required to generalize to the given test domain. Compared to *Debiasing*, the training set of *Domain Adaptation* is also class-balanced and there also exists an underlying bias in the training domain. The main difference is that unbiased data (testing data without labels) is available in the *Domain Adaptation*, indicating that it has more information for the unbiased data distribution than the *Debiasing* task. *Long-Tailed Recognition* [41–44], where the class-imbalanced training set is given and the deep model needs to be generalized to the class-balanced testing set. The main difference compared to *Debiasing* is that the bias is mainly caused by the imbalanced class distribution in the *Long-Tailed Recognition*, while there is no such case in the *Debiasing* due to its training classes are balanced *Class-Incremental Learning* [45–48] where a full training set with some base classes is given and the model needs to continuously learn new classes without keeping the original data. The bias in *Class-Incremental Learning* is caused by the imbalanced data distribution in the learned classes, which is similar to the *Debiasing*. However, it needs to continually remove the bias in each learning stage when class increments, which can be seen as a multi-stage *Debiaing* task. *Few-Shot Learning* [49–53], where a full training set with some base classes is given and the model needs to generalize to unseen classes in the testing set with the help of a small support set containing several image-label pairs. The bias in *Few-Shot Learning* is caused by the imbalanced data distribution in the base classes, and the difference compared to the *Debiaing* is that it needs to test on the new classes with few supports sample, while *Debiaing* task need the model to be tested on the original classes.

In this thesis, we focus on some specific cases for each camp: *Visual Dialog* for OOD Generalization with Multiple Modalities, which is a typical task to show the benefits of causal analysis for bias removal; *Domain Generalization* for OOD Generalization with Multiple Domains, which is the most commonly used task to study the OOD Generalization problem; *Debiasing* for OOD Generalization with no

Additional Annotations, which is the most challenging one, where the distribution shift is unlabelled (*e.g.*, different from Long-Tailed Recognition, where the shift of class distribution is known) and even unavailable (*e.g.*, different from Domain Adaptation, where the OOD data is available). We will discuss the related methods for them in detail in the following parts.

2.1.1 OOD Generalization with Multiple Modalities

In this camp, we mainly focus on the case Visual Dialog [24, 54], where the models are asked to give a good answer when given an image, as well as a dialog history of past question-answer pairs associated with the image, and the current question. It is more interactive and challenging than other vision-language tasks, such as Image Captioning [55–57], without the need for the language input and Visual Question Answering [27, 58, 59], which can be seen as a one-round Visual Dialog. In Specific, the first large-scale free-form Visual Dialog dataset, VisDial [24], applies a novel protocol to collect data: the questioner has to ask open-ended questions to an image without looking at it, while the answerer should give a free-form answer response to the question. There is another Visual Dialog dataset GuessWhat?! [54], which is a goal-driven dataset: the questioner has to locate an unknown object in an image by asking several closed-ended questions with only two answers "yes" and "no". We use the first setup because it is more challenging and more widely used. Therefore, causal analysis of Visual Dialog should consider the first format of the data collection process, in which the user plays an important role.

All of the traditional VisDial methods apply the typical encoder-decoder framework [60–65], which first encodes the question, dialog history and image into a joint embedding and then decode the answer by comparing to the answer candidates. By the usage of dialog history, we can categorize these methods into three camps: 1) Holistic history encoding. They treat dialog history as a whole text to feed into encoders, such as DAN [66], CoAtt [67], HACAN [26] and CorefNMN [68]. 2) Hierarchical history encoding. They apply a hierarchical structure to process dialog history, such as HRE [24]. 3) Recursive history encoding. They recursively encode the dialog history round by round. However, all of them overlook two facts, first, dialog history should not be directly fed to the answer model (*i.e.*, our proposed Principle 1), which will induce the history bias; Second, VisDial is essentially an

OOD task, where only one answer for the corresponding question is given in the dialog under the training stage but the models are required to treat all suitable answers as correct in the testing (See the evaluation details of VisDial in Section 3.6.1). And the models fail to generalize to test due to the user bias caused by the underlying confounders in the training (*i.e.*, our proposed Principle 2). Therefore, in Chapter 3, we propose two causal principles for Visual Dialog to help any traditional method to remove the bias to achieve better OOD performance.

2.1.2 OOD Generalization with Multiple Domains

In this camp, we focus on Domain Generalization, which tries to train a model on multiple training domains and test its generalization ability to unseen domains. To this end, all Domain Generalization methods aim to remove the domain-specific features and keep the domain-invariant/causal features. Depending on whether using the given domain annotations, they can be divided into two camps:

1) Without using domain annotations. They use a domain-agnostic augmentation/regularization to help the model learn more generalizable features [69–72]. For example, only learning features with small gradient is proposed by [69], due to they believe the easy-to-learn features are more likely the biased features; a regularization for class features is proposed and the effective mixup augmentation is applied in [72]. However, DomainBed [12] shows that a strong Empirical Risk Minimization (ERM) Learning based model beats most of them, implying that the improvements claimed by these methods are mostly due to an unfair hyperparameter tuning, demonstrating that they fail to remove the domain bias. Note that, Domain Generalization without domain labels can be treated as Implicit Debiasing, which we will introduce in Section 2.1.3 and Chapter 5. Therefore, we focus on the other camp in Chapter 4.

2) Using domain annotations. They use domain-wised regularization to encourage models to eliminate domain-specific bias, and then the domain-invariant features are achieved to realize generalization. In specific, they include:

Invariant/causal learning [4, 73–75], where the simultaneously optimal in each environment is encouraged by a designed invariance penalty, such as gradient penalty

of a dummy classifier [4], or Euclidean distance of the loss between different environments [73], or to optimize the worst group to realize the equal optimal is proposed by [74].

Feature/gradient alignment [32, 76–79], where the distance of class features [76] or gradients of the same class [77] from different domains are minimized. Sometimes, the distance of the higher degree of moments of a class between different domains is also regularized to achieve the domain invariance [32].

Adversarial learning [31, 80], where the domain labels are explicitly learned by an additional head to generate adversarial gradients to the feature extractor, which let the model cannot predict the domain labels, *i.e.*, cannot extract the domain-specific features. Therefore, the domain-specific bias will not be learned by the model.

However, in Chapter 4, we will show that all the current methods suffer from spurious invariance, which cannot eliminate all the bias in different domains, and our proposed DOMAIN+ will address the problem to improve OOD performance.

2.1.3 OOD Generalization with no Additional Annotations

In this camp, we mainly focus on (Implicit) Debiasing, where the methods attempt to train an unbiased model on a biased training set, and there are no additional annotations, compared to Explicit Debiasing, which requires the pre-defined bias and its annotation. Note that, in practice, it is impossible to know all biases and have their full annotations, so Implicit Debiasing is more often studied.

We can view all current implicit debiasing methods as one-stage re-weighting methods. In general, they train a conventional model on biased data and treat it as the bias model, whose false positive confidence can be seen as the bias level, which is proportional to the weight of the samples in the re-weighting, *i.e.*, the higher the false positive confidence, the more likely the sample is to be biased in the training, and the higher weight should be assigned to the sample in the unbiased training. In addition to the vanilla method [2] which combines training the bias model, bias estimation and re-weighting in the same stage, we have the following variants:

Re-weighting with pre-estimated bias [81], where the bias model is learned and the bias is estimated first and then they use the bias to re-weighted train the unbiased model from scratch. Although they apply two stages, it is still a one-stage re-weighting method because the bias estimation stage can be replaced by any bias estimation strategy.

Re-weighting by regularization [82], where the bias model is learned first and then they impose a regularization loss to emphasize samples whose predictions are independent of the bias model, where such independence can be considered as unbiasedness.

Re-weighting with augmentation, where they follow the training framework of LfF [2] and additional feature-level augmentation is applied by randomly concatenating biased features to unbiased features in the re-weighted training [6], or performing sample-level augmentation with samples generated by Variational Autoencoders (VAE) which is pretrained on the higher-weight samples [83].

Re-weighting with the supervised contrastive loss [84], where they apply the supervised contrastive loss [85] to regularize the higher-weight samples, and it can be viewed as the one-stage implementation of our method in Section 5.2.

As we will discuss in Section 5.2, these one-stage methods suffer from the "anti-training bias" bias, which leaves their "unbiased" models still biased, and this issue will be specifically addressed by our method. Interestingly, our two-stage design coincides with the popular "biased training and unbiased adjustment" approach in long-tail classification [43, 86, 87]. Although we differ in the type of bias (our attribute bias vs. their class size bias) and the second stage of training (they only fine-tune/tune the head of the classifier by using pre-calculated weights on class size bias), we believe that a unified theory may exist.

Furthermore, due to the mainly used re-weighting in current debiasing methods [2, 6, 82] is based on incorrect context estimation, which wrongly estimates context by the entanglements of the class effect and context effect, in Section 5.3, we propose a better context estimation by disentagling context features to achieve better re-wighting for removing bias for OOD Generalization.

2.2 Related Technologies

In our implementations for removing the various types of bias in different OOD Generalization tasks, we are motivated by causal inference [19] and inspired by some specific techniques, such as influence function [88] and invariant feature learning strategies [4, 73]. Therefore, we introduce these related technologies in this section.

2.2.1 Causal Inference

Causal Inference tries to explore the data generation process by pursuing the causal relationship between variables to realize robust inference. Some works [89–92] introduce causal theory into machine learning, attempting to endow the model with the abilities of causal reasoning in the learning process. In contrast to them, we focus on the specific cases and explicitly use causal graph [19] to reveal the nature of the data and do some causal analysis to guide our training.

Causal Inference with the causal graph is also related to some works, which can be categorized as deconfounding [93–95] and counterfactual inference [43, 96, 97]. Specifically, a learnable dictionary is built to deconfound the invisible confounders in vision-language pretraining [93] and semi-supervised detection [94]; the counterfactual effect analysis, such as total direct effect and natural indirect effect, based on the causal graph, is implemented in long-tailed classification [43] and visual question answering [97].

2.2.2 Influence Function

The Influence Function is rooted in robust statistics [98], which approximates the impact on certain objective functions when perturbing a sample from the training data by an infinitesimal value and re-train the model from scratch. The larger sample influence denotes the sample is more harmful to the objective during training. Recent works use the influence function to measure the sample or feature influence on testing loss [88, 99–102]. For example, sample-wise influence on the testing loss is estimated in [88], and then they remove the harmful samples, that have a higher influence on the loss of the testing set, and re-train the model to get lower loss in the testing set; feature-wise influence is estimated in [100], and then they apply

the re-weighting method for features to eliminate the harmful features; relabelled sample-wise influence is estimated to utilize the harmful samples by changing their labels rather than removing them in the training to decrease the testing loss is proposed by [101]. Therefore, we can view them as oracle tuning tricks due to they need the testing data to calculate the influence, and thus their application is limited due to the testing distribution is not always available. To solve the limitation, in Chapter 4, we propose a cross-domain influence that only requires training data to find the "harmful" samples. In particular, our influence calculation is based on the efficient estimation method in [88].

2.2.3 Invariant Feature Learning

In image classification, the invariant features denote the class/causal features that can discriminate the class under any data distribution. Therefore, the bias is removed and robust classification is achieved as long as the invariant features are learned by the model. The prevailing methods are:

Data augmentation [5, 7, 8, 103], where some augmentations are pre-defined for images to enlarge the distribution of class-irrelevant features artificially and then let the model learn invariance to the augmentations. Specifically, the jigsaw puzzles are proposed in [5], which lets the model solve the puzzle game of image pieces as an additional objective; the adversarial augmentation on the image is proposed in [7] and mixup augmentation is proposed in [103]. However, these augmentation-based methods only make the model invariance to the augmentation-related bias and as it cannot cover all class-irrelevant bias or some pre-defined augmentation are ill-posed, the incomplete and inaccurate bias-removing will impact their feature invariance.

Causal Learning [16, 104–106]. They learn causal representations to capture the underlying data generation process, and then they can theoretically eliminate bias features and keep causal features to pursue the causal effect by intervention. However, these methods finally are essentially implemented as re-weighting methods in the causal view. Note that this subsection can be seen as the implementation of Causal Inference we have introduced in Section 2.2.1.

Re-weighting [2, 6, 9]. Their methods are based on the Inverse Probability Weighting [19, 107, 108] to use sample-wise weight to rebalance the bias distribution, helping the model to learn the invariance. Specifically, a bias model is trained by Generalized Cross Entropy loss [109] to emphasize the bias learning, and then it is used to assign weights for samples based on the estimated bias distribution [2]; a further feature permutation-based augmentation is proposed by [6], which improves the performance of re-weighting.

Chapter 3

OOD Generalization with Multiple Modalities[1]

3.1 Case Study: Visual Dialog

Given an image I, a dialog history H of pairs of the past question Q and answer A (Q&A): $H = \{(Q_1, A_1), ..., (Q_{t-1}, A_{t-1})\}$, and the current t-th round question Q, a Visual Dialog (VisDial) agent [24] is expected to give a good answer A. Our community has always considered VQA [27] and VisDial as sister tasks due to their similar settings: Q&A grounded by I (VQA) and Q&A grounded by (I, H) (VisDial). Indeed, from a technical view, just like the VQA models, a typical VisDial model first uses an encoder to represent I, H, and Q as vectors, and then feeds them into a decoder for answer A. Thanks to the recent advances in encoder-decoder frameworks for VQA [58, 110], as well as for natural language processing [111], the performance (NDCG [112]) of VisDial in literature is significantly improved from the baseline 51.63% [10] to the state-of-the-art 64.47% [63].

However, we would like to highlight an important fact that VisDial is essentially not VQA. And this fact is so profound that all the common heuristics in the vision-language community, such as the fusion tricks [58, 113] and attention variants [110, 114], cannot appreciate the difference. Therefore, all current VQA-based methods overlook the new bias introduced by the VisDial framework compared to VQA.

[1]The main content in this chapter is published as **Jiaxin Qi**, Yulei Niu, Jianqiang Huang, and Hanwang Zhang. Two Causal Principles for Improving Visual Dialog. *CVPR 2020*

Instead, we use the *causal inference* [19, 115]: a graphical framework that stands in the cause-effect of the data, but not merely the statistical association of them, to explore the specific bias in VisDial and propose two causal principles:

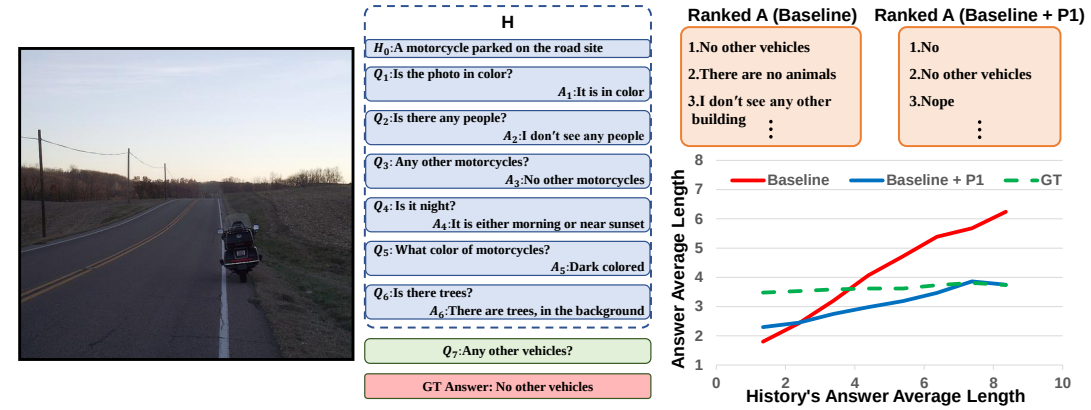
Principle 1. (**P1**): We should delete the direct link $H \rightarrow A$.

Principle 2. (**P2**): We need to add one new node (unobserved) U and three new links: $U \leftarrow H$, $U \rightarrow Q$, and $U \rightarrow A$.

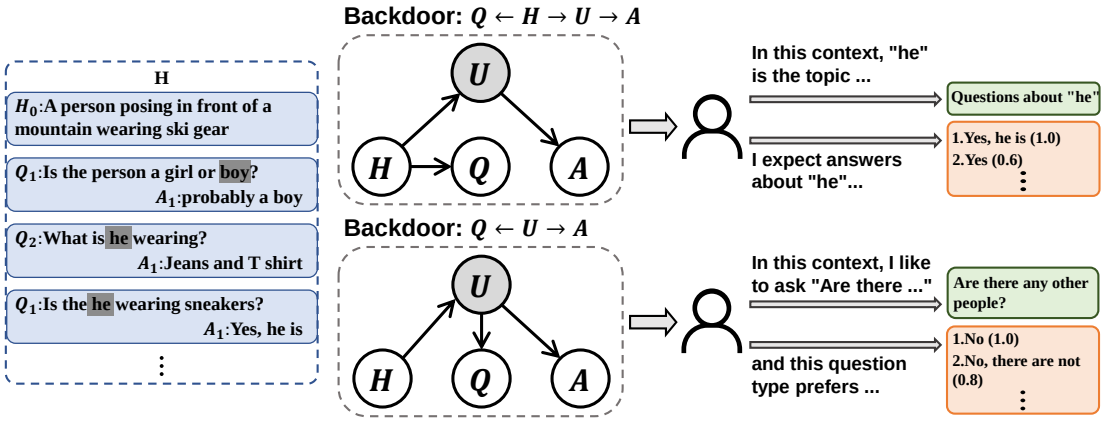
Figure 3.3 compares the causal graphs of existing VisDial models and the one applied with the proposed two principles. Although a formal introduction of them is given in Section 3.3, now you can simply understand the nodes as data types and the directed links as data flows. For example, $V \rightarrow A$ and $Q \rightarrow A$ indicate that the visual knowledge V, *e.g.*, the encoded feature from a multi-model encoder, works with the question Q to "dictate" the answer A.

P1 suggests that we should remove the direct input of dialog history to the answer model to remove the history bias. This principle contradicts most prevailing VisDial models [24–26, 60, 63, 64, 66, 67], which are based on the widely accepted intuition: the more features you input, the more effective the model is. It is mostly correct, but only with our discretion in the data generation process. In fact, the annotators of the VisDial dataset [24] were not allowed to copy from the previous Q&A, *i.e.*, $H \nrightarrow A$, and were encouraged to ask consecutive questions that include co-referenced pronouns like "it" and "those", *i.e.*, $H \rightarrow Q$, and the answer A should be based only on question Q and the reasoned visual knowledge V. Therefore, a good VisDial model is expected to reason over the context (I, H) with Q but not to memorize the bias. However, the direct path $H \rightarrow A$ will contaminate the expected causality. Figure 3.1(a) shows a very ridiculous bias observed in all baselines without P1: the top answers are those whose lengths are close to the average length in the history answers.

P2 implies that the model training based only on the association among the sample (I, H, Q) and A is spurious. By "spurious", we mean that the effect on A caused by (I, H, Q), the goal of VisDial, is *confounded* by an unobserved variable U, because it appears in every undesired causal path (*a.k.a.*, backdoor [19]), which is an indirect causal path from the input (I, H, Q) to output A: $Q \leftarrow U \rightarrow A$ and $Q \leftarrow H \rightarrow U \rightarrow A$. We believe that such unobserved U should be *users* as the VisDial dataset essentially brings humans in the loop. Figure 3.1(b) illustrates



(a) A Typical Dialog History Bias



(b) User Preference Bias

FIGURE 3.1: The illustrative motivations of the two causal principles: Principle 1, we should cut the direct link between dialog history H and answer A, which introduces history bias such as the example shown in (a) and Principle 2, We need to add one new node (unobserved) U and three new links: $U \leftarrow H$, $U \rightarrow Q$, and $U \rightarrow A$ and de-confound the confounder U, which introduces user preference bias as shown in (b). Refer to Section 3.3 and Figure 3.3.2 for more details of the causal graph.

how the user's hidden preference confounds them, as the VisDial dataset essentially involves humans in the loop. Therefore, during training, if we focus only on the conventional likelihood $P(A|I, H, Q)$, the model will inevitably be biased towards the spurious causality, *e.g.*, it may score the answer "Yes, he is" higher than "Yes", merely because the users prefer to see a "he" appeared in the answer, given the history context of "he". It is worth noting that the confounder U is more impactful in VisDial than in VQA, because the former encourages the user to rank similar answers subjectively while the latter is more objective. A plausible explanation might be: VisDial is interactive in nature and a not quite-correct answer is tolerable

in one iteration (*i.e.*, dense predictions); while VQA has only one chance, which demands accuracy (*i.e.*, one-hot prediction).

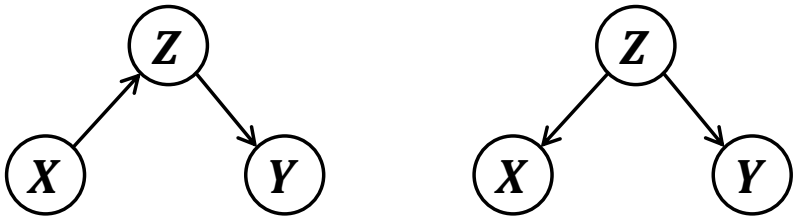
By applying P1 and P2 to the baseline causal graph, we have the proposed one (the right one in Figure 3.3), which serves as a *model-agnostic* roadmap for the causal inference of VisDial for removing the bias. To remove the bias caused by U, we use the *do*-calculus [19], $P\left(A|do(I,H,Q)\right)$, which is fundamentally different from the conventional likelihood $P(A|I,H,Q)$, where the former is an active intervention, which cuts off $U \rightarrow Q$ and $H \rightarrow Q$, and sample every possible $U|H$, seeking the true effect on A only caused by (I,H,Q); while the latter likelihood is a passive observation that is affected by the existence of U. The formal introduction and details will be given in Section 3.4.3. In particular, given the fact that once the dataset is ready, U is no longer observed, we propose a series of effective approximations in Section 3.5.

Before we delve into the details, we would like to summarize the main contributions: two causal principles, derived from the analysis of the difference between VisDial and VQA, which lead to a performance leap — a farewell to the 60%-s and an embrace of the 70%-s — for all the VisDial models.[2] Specifically, we apply our principles for removing the bias to improve four baseline models: LF [24] (↑16.42%), HCIAE [116] (↑15.01%), CoAtt [67] (↑15.41%), and RvA [25] (↑16.14%). Impressively, on the official test-std server, we use an ensemble model of the most simple baseline LF [24] to beat the 2019 champion (which is also our methods with P1 and P2) by 0.2%, a more complex ensemble to beat it by 0.9%, and lead all the single-model baselines to the state-of-the-art performance.

3.2 Preliminaries: Causal Theory

Due to the training data of Visual Dialog involving multiple modalities, we need to take a causal perspective to analyze the nature of the data generation process to find the bias. In order to describe causal relationships between different variables, we need to first introduce a graphical framework, the causal graph, designed by Judea Pearl [19] and then briefly introduce some basic causal preliminaries, including

[2]Only those with codes&reproducible results due to resource limit.



(a) Chain (b) Fork

FIGURE 3.2: Two basic causal graph structures: (1) Chain, (2) Fork.

confounder and *do*-calculus. In this subsection, we use X, Y, Z to denote variables and x, y, z to denote their values.

3.2.1 Causal Graph

Causal graph [19] is a high-level roadmap, which indicates the causal interactions between variables, to reveal the underlying causal relationships. Generally, we use a directed acyclic graph to illustrate it. A causal graph includes nodes, which denote variables, and directed links, which denote causal relationships. For example, in Figure 3.2, the left graph illustrates one of the basic causal graph structures, contains $X \to Z$, denoting that X causes Z, and $X \not\to Y$, denoting there is no causal effect from X to Y.

There are two basic structures for three nodes in the causal graph: Chain and Fork, which are illustrated in Figure 3.2. The Chain describes the structure that, for the middle node Z, there is a directed link, $X \to Z$, points into it and a directed link points out of it, $Z \to Y$. Fork describes the structure that, for the middle node Z, there are two directed links emanating from it, $Y \leftarrow Z \to X$.

There are several dependencies in these structures: 1) X and Z are dependent in Figure 3.2(a), as $P(Z = z|X = x) \neq P(Z = z)$ (two variables with a directed link are dependent); 2) X and Y are independent, conditional on Z, in Figure 3.2(b), as $P(Y = y|Z = z, X = x) = P(Y = y|Z = z)$ (two variables without a directed link can be conditional independent). It is recommended to read the book [19] for more details. Here, we use the classical "ice cream and crime" example to illustrate the conditional independence: We use X, Y and Z to denote "sales of ice cream", "crime ratio" and "temperature" in Figure 3.2(b), respectively. The

directed links denote the data generation logic (causal relationships) between them, such as the increase of "temperature" Z causes the increase of "sales of ice cream" X, *i.e.*, $Z \rightarrow X$; the increase of "temperature" Z causes the increase of "crime ratio" Y (*e.g.*, harassment in summer), *i.e.*, $Z \rightarrow Y$. If we condition on Z, *i.e.*, set Z equal to some certain value z, such as observing the relationship between X and Y in summer, we will find that there is no correlation between them (conditional independence between X and Y).

3.2.2 Causal Effect

In the naive causal graph $X \rightarrow Y$, with no other nodes and arrows. The effect of X on Y should be $P(Y|X) - P(Y)$, where $P(Y)$ is a constant prior (we sometimes use $P(Y|X)$ to represent the effect of X on Y for convenience). Because there is only one causal path from X to Y, the effect can only pass through the causal path. So, $P(Y|X)$ is the causal effect that we want to pursue. However, in the real world, things are not easy like this.

3.2.3 Confounder

As shown in Figure 3.2(b), the variable Z in the fork structure is called **confounder**, which opens the **backdoors** (paths start from the confounder and ends at X and Y or their descendants) and will introduce the spurious correlations between X and Y, makes the model to learn the bias. We use the previous example to further demonstrate the statement. In our commonsense, the "sales of ice cream" cannot cause the "crime ratio", *i.e.*, we cannot use the "sales of ice cream" to predict the "crime ratio". For example, if one day, the "sales of ice cream" increase because of the discount, we cannot predict that the "crime ratio" will increase on that day. However, if we train a model based on the data of X and Y by $P(Y|X)$, we will find that X and Y have a positive correlation because they both increase in summer and decrease in winter (they have statistical correlations). The result is that the model will wrongly predict Y by X, based on their co-occurrence with the confounder Z, which we called $P(Y|X)$ is biased by Z.

The underlying reason for deep models learning $P(Y|X)$ is biased because $P(Y|X)$ contains spurious correlations and does not describe the causal effect from X to Y

when there is a confounder between them. In the "ice cream and crime" example, the model learned from the observational data will predict that $P(Y|X) - P(Y) \neq 0$ as we discussed in the previous section, and thus the learned effect from X to Y by deep models is different from the causal effect, 0, *i.e.*, "sales of ice cream" X has no causal effect on the "crime ratio" Y. To remove the learned bias induced by the confounder, we need some causal tools to pursue the causal effect between X and Y, instead of learning $P(Y|X)$.

3.2.4 *do*-calculus

Judea Pearl [19] introduces the *do* factor to describe the causal effect from X to Y, $P(Y|do(X = x)) - P(Y)$, where $P(Y)$ is the prior and can be omitted as we mentioned. $do(X = x)$ denotes that, instead of observing $X = x$ from its distribution, we intervene X equal to a certain x and then observe the following results, like the randomized independent experiment, in the following, we use $do(X)$ to denote $do(X = x)$ for convenience. For example, in the previous "ice cream and crime" example, $P(Y|do(X))$ denotes we actively increase or decrease the price to intervene X equal to a certain value and then observe the "crime ratio" on that day. Informally, we will find that no matter how we set X equal to a certain value x, the result of the "crime ratio" will not change, indicating there is no effect from X to Y, which is actually the causal effect. Therefore, we can use $P(Y|do(X))$ to denote the causal effect. We provide the formal justification in the following.

When the data for X and Y are given, intervening X and observing the results becomes impossible. Thus, we need to calculate the *do* formula by the observational data, *i.e.*, transform the *do* formula into the probability formula. Thanks to the *do*-calculus in the book [19], there are three basic rules to help us remove *do*.

Rule 1. If X is not the cause of Y, *i.e.*, no causal path starts from X and ends at Y, we can simply remove $do(X)$:

$$P(Y|do(X)) = P(Y). \tag{3.1}$$

Rule 2. When X and Y are conditionally independent (conditional on Z), we can simply remove X:

$$P(Y|z, X) = P(Y|z). \tag{3.2}$$

Rule 3. If all the backdoors from X to Y are blocked by Z, *e.g.*, conditioning on Z is to block the backdoors through Z, $do(X) = observe(x)$:

$$P(Y|do(X), z) = P(Y|x, z). \tag{3.3}$$

Based on the three *do*-calculus rules and the causal graph in Figure 3.2(b), we can calculate the formula $P(Y|do(X)) - P(Y)$ in the "ice cream and crime" example:

$$\begin{aligned} &P(Y|do(X)) - P(Y) \\ &= \sum\nolimits_z P(Y|do(X), z)P(z|do(X)) - P(Y) \\ &= \sum\nolimits_z P(Y|X, z)P(z|do(X)) - P(Y) \\ &= \sum\nolimits_z P(Y|X, z)P(z) - P(Y) \\ &= \sum\nolimits_z P(Y|z)P(z) - P(Y) \\ &= P(Y) - P(Y) = 0, \end{aligned} \tag{3.4}$$

where the first derivation is using the Bayes Rules, and the second derivation is using **Rule 3**, where Z blocks all backdoors from X to Y, *i.e.*, $P(Y|do(X), z) = P(Y|X, z)$, and the third derivation is using **Rule 1**, where X is not the cause of Z, *i.e.*, $P(z|do(X)) = P(z)$, and the fourth derivation is using **Rule 2**, where X and Y are conditionally independent (conditioning on Z), *i.e.*, $P(Y|X, z)P(z) = P(Y|z)$. We find that the effect calculated by $P(Y|do(X)) - P(Y)$ is 0, which again conforms to the true causal effect 0 in the "ice cream and crime" example. Therefore, $P(Y|do(X))$ describes the causal effect, and the model trained by $P(Y|do(X))$ can remove the bias induced by the confounder.

3.2.5 De-confounding and Re-weighting Implementation

In this subsection, we would like to justify the underlying causalities of the popular implementation, re-weighting, which is the prototype of most of our proposed methods. To achieve the unbiased causal effect $P(Y|do(X))$ under the confounder Z, we can adjust the traditional objective $P(Y|X)$ by a sample-wise weight (just

like the Eq (3.4)):

$$\begin{aligned} P(Y|do(X)) &= \sum_z P(Y|do(X), z)P(z|do(X)) \\ &= \sum_z P(Y|X, z)P(z|X) \\ &= \sum_z P(Y|X, z)P(z) \\ &= \sum_z \frac{P(Y, X, z)}{P(X|z)P(z)}P(z) \\ &= \sum_z P(Y, X, z)\frac{1}{P(X|z)}, \end{aligned} \tag{3.5}$$

where the first two derivations are similar to the Eq (3.4), and the last two derivations are based on the Bayes Rules, where $P(X|z)$ is also called as propensity score [19, 107, 108] and $P(Y, X, z)$ can be estimated by the traditional objective function:

$$\begin{aligned} P(Y|X) &= \sum_z P(Y|X, z)P(z|X) \\ &= \sum_z \frac{P(Y, X, z)P(z, X)}{P(X, z)P(X)} \\ &= \sum_z P(Y, X, z)\frac{1}{P(X)}, \end{aligned} \tag{3.6}$$

where $P(X)$ is a prior. Therefore, by adding a weight $\frac{P(X)}{P(X|z)}$, which is also called inverse probability weighting [117, 118], we can calculate the unbiased effect $P(Y|do(X))$ by using the traditional objective $P(Y|X)$, which is the causal justification for why re-weighting is the implementation for achieving the causal (unbiased) effect.

3.3 Visual Dialog in Causal Graph

In this section, we first formally introduce the Visual Dialog (VisDial) task and then describe how the popular encoder-decoder framework follows the baseline causal graph shown in Figure 3.3, from which we can summarize the causal principles to be used in Section 3.4 to remove the bias.

3.3.1 Visual Dialog Settings

Settings. According to the definition of VisDial task proposed by Das [24], at each time t, given an input image I, current question Q_t and dialog history $H = (C, (Q_1, A_1), \ldots, (Q_{t-1}, A_{t-1})$ about the image, where C is the image caption, (Q_i, A_i) is the i-th round question-answer pair, and a list of 100 candidate answers $A_t = \{A_t^1, \ldots, A_t^{100}\}$ for deterministic models to choose from, the task of the dialog agent is to generate a free-form answer or give an answer by ranking candidate answers A_t to answer the current question Q_t according to the image and dialog history.

Evaluation. Recently, the VisDial community has adopted the ranking metric Normalised Discounted Cumulative Gain (NDCG). It differs from the classification metric (*e.g.*, top-1 accuracy) used in VQA. It is more compatible with the relevance scores of answer candidates evaluated by humans in VisDial. NDCG requires that relevant answer candidates be ranked higher than just the selection of ground truth answers, and more details of NDCG can be found in [112].

3.3.2 Traditional Visual Dialog Causal Graph

Based on the basics of causal graphs that we introduced in Section 3.2.1, we revisit the popular VisDial encoder-decoder framework in existing methods using elements of the baseline graph in Figure 3.3.

Feature Representation and Attention in Encoder. Visual feature is denoted as node I in the baseline graph, which is usually a fixed feature extracted by Faster-RCNN [119] based on ResNet backbone [120] pre-trained on Visual Genome [121]. For language feature, the encoder firstly embeds sentence into word vectors, followed by passing the RNN [122, 123] to generate features of question and history, which are denoted as $\{Q, H\}$.

Most existing methods apply the attention mechanism [124] in the encoder-decoder framework to explore the latent weights for a set of features. A basic attention operation can be represented as $\tilde{\boldsymbol{x}} = Att(\boldsymbol{x}, \boldsymbol{k})$ where $\boldsymbol{x}$ is the set of features need to attend, $\boldsymbol{k}$ is the key (*i.e.*, guidance) and $\tilde{\boldsymbol{x}}$ is the attended feature of $\boldsymbol{x}$. Details can be found in most visual dialog methods [26, 67, 116]. In the baseline graph

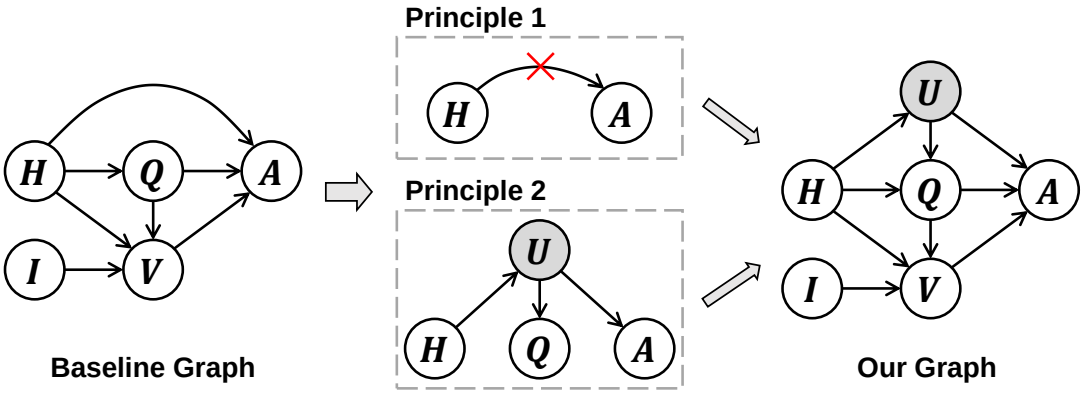


FIGURE 3.3: Causal graphs of VisDial models (baseline and ours). H: dialog history. I: image. Q: question. V: visual knowledge. A: answer. U: user preference. Shaded U denotes the unobserved confounder. See Section 3.3.2 for detailed definitions.

illustrated in Figure 3.3, the sub-graph $\{I \rightarrow V, Q \rightarrow V, H \rightarrow Q \rightarrow V\}$ denotes a series of attention operations for visual knowledge V. Note that these arrows are not necessarily independent, such as co-attention [67], and the process can be written as $Input : \{I, Q, H\} \Rightarrow Output : \{V\}$, where intermediate variables can be yielded in the graph with respect to different attention strategies such as co-attention [67] and recursive attention [25]. However, without loss of generality, these variables do not affect the causalities in the graph.

Response Generation in Decoder. After obtaining the features from the encoder, existing methods will fuse these features and feed the fused one into a decoder to generate an answer. In the baseline graph, node A denotes the answer sentence and the generation process is that the decoder takes the features via $\{H \rightarrow A, Q \rightarrow A, V \rightarrow A\}$ and transforms them into a vector for decoding the answer. In particular, the decoder can be generative, *i.e.*, to generate an answer sentence by RNN; or discriminative, *i.e.*, select an answer by discriminating answer candidates.

Next, we advance to the middle part of Figure 3.3, presenting two causal principles that reveal the bias that existed in the traditional VisDial methods.

3.4 Two Causal Principles

3.4.1 Principle 1

P1: We should delete the direct link $H \rightarrow A$. When should we draw an arrow from one node pointing to another? According to the definition in Section 3.2.1, the criterion is if the node causes another one. Intrigued, let's understand P1 by discussing the rationale behind the "double-blind" review policy. Given three variables: "Well-known Researcher" (R), "High-quality Paper" (P), and "Accept" (A). From our community common sense, we know that $R \rightarrow P$ because top researchers usually lead high-quality research, and $P \rightarrow A$ is even more obvious. Therefore, for the good of the community, the double-blind prohibits the direct link $R \rightarrow A$ by author anonymity, otherwise, the bias such as personal emotions and politics from R may affect the outcome of A.

The story is similar in VisDial. Without loss of generality, we only analyze the path $H \rightarrow Q \rightarrow A$. If we inspect the role of H, we can find that it is to help Q resolve some co-reference like "it" and "their". As a result, H is the cause of Q. Then, A is given according to Q by the answerer. Here, Q becomes a mediator which cuts off the direct association between H and A that makes $P(A|Q,H) = P(A|Q)$, like the "High-quality Paper" which we mentioned in the previous story. However, if we draw an arrow from H to A to describe the causality: $H \rightarrow A$, the undesirable bias of H will be learned for the prediction of A, that hampers the natural process of VisDial, which is the history bias and an example is illustrated in Figure 3.1(a). Another example is discussed in Figure 3.5 that A prefers to match the words in H even though they are literally nonsense about Q if we add the direct link $H \rightarrow A$. After we apply P1, these phenomena will be relieved, such as the blue line illustrated in Figure 3.1(a), which is closer to the NDCG ground truth (*i.e.*, candidates with non-zero relevance score) average answer length represented as green dash line, and the other qualitative studies in experiments.

3.4.2 Principle 2

P2: We need to add one new node (unobserved) U and three new links: $U \leftarrow H$, $U \rightarrow Q$, and $U \rightarrow A$. As we introduced in Section 3.2.3, in causal

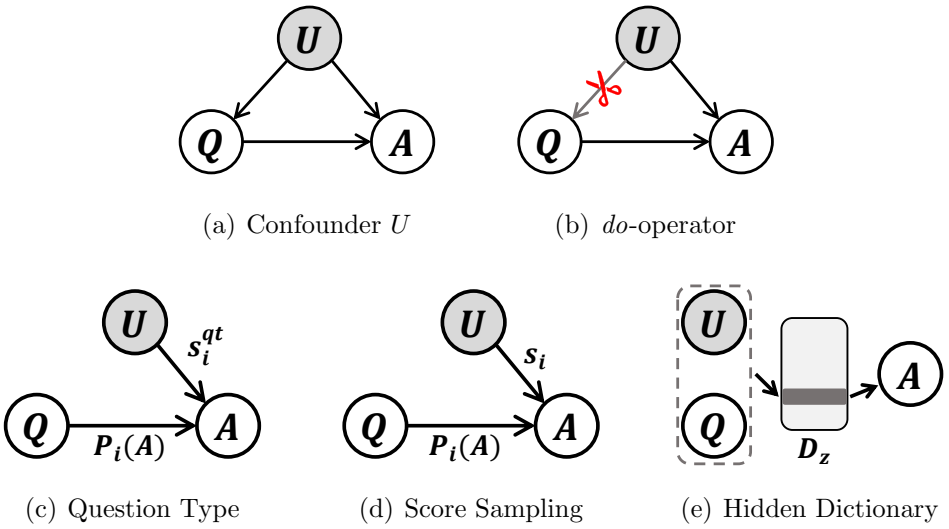
graph, the fork-like pattern in Figure 3.4(a) contains a *confounder* U, which is the common cause for Q and A (*i.e.*, $Q \leftarrow U \rightarrow A$). The confounder U opens a non-causal path started from Q which is also called the *backdoor*, making Q and A spuriously correlated even if there is no direct causality between them. To remove the bias induced by the confounder U, we need to de-confound U.

In the data generation process of VisDial, we know that not only both the questioner and answerer can see the dialog history which offers them a latent topic, but also the answer annotators can look at the history when annotating the answer. Their preference can be understood as part of human nature or subtleties conditional on a dialog context, and thus it has a causal effect on both Q and A. Moreover, due to the fact that the preference is nuanced and uncontrollable, we consider it as an *unobserved* confounder for Q and A.

It is worth noting that the confounder hinders us to find the true causal effect. Let's take the graph in Figure 3.4(a) as an example, if there is no U, the probability $P(A|Q)$ is the causal effect that we want to pursue. However, due to the existence of U, $P(A|Q)$ is no longer the true causality from Q to A. When we calculate $P(A|Q)$, we take U into account which can be shown by Bayes rule:

$$P(A|Q) = \sum\nolimits_u P(A|Q,u)P(u|Q). \tag{3.7}$$

The distribution of u is conditional on Q (*i.e.*, $P(u|Q)$). That means when using the conditional weight (*i.e.*, $P(u|Q)$) to sum every effect (*i.e.*, $P(A|Q,u)$), the likelihood sum (*i.e.*, $P(A|Q)$) will be biased towards the effect $P(A|Q,u)$ with larger weights. For better understanding, if we treat Eq (3.7) as a process of data stratification, at each layer u, we can obtain the causality conditional on u, because given u will block the backdoor of Q. Then, we have to sum these causalities by the natural distribution of u rather than conditional distribution $P(u|Q)$, which will remix the data bias. In a nutshell, we cannot calculate causality from Q to A by $P(A|Q)$ under the confounder U. To resolve this problem (*i.e.*, de-confounding to find causal effect), we need more powerful tools.



(a) Confounder U (b) *do*-operator

(c) Question Type (d) Score Sampling (e) Hidden Dictionary

FIGURE 3.4: Example of (a) confounder U, (b) *do*-operator and (c)-(e) sketch causal graphs of our three attempts to de-confound

3.4.3 Our Visual Dialog Causal Graph

As we introduced in Section 3.2.4, we can use the *do*-operator and *do*-calculus [19, 125] to help us de-confound.

***do*-operator.** *do*-operator is a type of intervention to de-confound. Illustrated in Figure 3.4(b), *do*-operator (*e.g.*, $do(Q = q)$) is that we set a value q to variable Q, *i.e.*, Q is caused by itself rather than its parent nodes. Therefore, $do(Q = q)$ cut off all the original arrows that come into Q (*i.e.*, $U \rightarrow Q$) because its parents do not cause it anymore. This operation can prevent any information about Q from flowing in the non-causal direction (*i.e.*, backdoor $Q \leftarrow U \rightarrow A$). As a result, the confounder of Q can be relieved and the causal effect of Q can be estimated. In the following parts, we use $do(q)$ to represent $do(Q = q)$ for concision.

***do*-calculus.** As we discussed, it is hard to take a real intervention on a fixed dataset. We need to use the *do*-calculus to translate $P(A|do(q))$ into $P(A|(Q, \dots))$, where the later one has no *do*-operator and can be calculated by conditional probability. The rules of *do*-calculus are given in [19, 125] and here we review the most important one: If a set Z of variables blocks all backdoor paths from X to Y, then conditional on Z, $do(x)$ is equivalent to $observe(x)$: $P(Y|do(x), Z) = P(Y|X, Z)$. Other rules are given in Section 3.2.4.

Now, we can revisit the example in Section 3.4.2. If we calculate $P(A|do(q))$ rather than $P(A|Q)$, the result will be $\sum_u P(A|Q,u)P(u)$. In this formula, the distribution of u is the natural prior $P(u)$ instead of the conditional distribution $P(u|Q)$. Therefore, the summation of the causal effect by weight (*i.e.*, $P(u)$) will not remix the data bias, *i.e.*, the confounding bias is removed by $P(A|do(q))$.

In our graph of VisDial shown in Figure 3.3, we can also de-confound U by intervention $do(q,h,i)$ to find causal effects from $\{Q,H,I\}$ to A, then perform do-calculus rules to transform pretended intervention into probability formula:

$$\begin{aligned} &P(A|do(q,h,i)) \\ &= \sum\nolimits_u P(A|do(q,h,i),u)P(u|do(q,h,i)) \\ &= \sum\nolimits_u P(A|do(q),H,I,u)P(u|H) \\ &= \sum\nolimits_u P(A|Q,H,I,u)P(u|H). \end{aligned} \tag{3.8}$$

The last transformation takes the rule we introduced in do-calculus because Q's backdoors are blocked by controlling U. Referring to the example in Section 3.2.4, the rest derivations are easy to prove. As we mentioned, $P(A|do(q,h,i))$ in our causal graph is the causal effect that we want in VisDial, which removes all bias.

So far, we have given all contents about the baseline causal graph, two principles, and our causal graph. In the next section, we will try to realize our causal graph by providing some implementation attempts to en-light the future of VisDial.

3.5 Removing Bias by Our Causal Graph

Removing history bias by P1. It is easy to implement P1 in the training, *e.g.*, we can simply cut the direct feeding of history into the answer generation process of the decoder. We will give some examples as training details in Section 3.6.3.

Removing user bias by P2. As for P2, we can obtain causal effect estimation by Eq (3.8) which can be written as:

$$P(A|do(q,h,i)) = \sum\nolimits_u P_u(A)P(u|H), \tag{3.9}$$

where $P_u(A)$ represents the probability of A under the conditions Q,H,I and u. Since the variable U is unobserved, we just give some examples of attempts to

replace U or approximate it and corresponding sketch graphs will be given to help understand in Figure 3.4.

1) Question Type. Inspired by the data stratification form in Eq (3.9), we try to use question type to stratify the data. Specifically, we manually define some question types, count appeared answers and, set preference for every answer in each type of question. According to the Eq (3.9), we can use the preference generated by question type to train our model with the loss function:

$$\mathcal{L}_{qt} = \sum\nolimits_i P_i(A) \cdot s_i^{qt}, \tag{3.10}$$

where i is the i-th candidate in the answer list, $P_i(A)$ is the probability of candidate i, s_i^{qt} is the preference we counted and the sketch graph is shown in Figure 3.4(c), and the implementation details will be given in Section 3.6.3.

2) Answer Score Sampling. The official gives a set of dense annotations (the relevance score for each candidate) in the training set which can be treated as a representation of preference because the annotators score every candidate in the context H with their preference. As a result, if we regard each candidate A_i in the decoder as a u, illustrated in Figure 3.4(d), we can follow Eq (3.9) to calculate loss by the following function:

$$\mathcal{L} = -\sum\nolimits_i P_i(A) \cdot s_i, \tag{3.11}$$

where i is the index of the answer candidate. Eq (3.11) can be implemented in different forms. Here we give three examples:

Weighted Softmax Loss (R_1). We extend the log-softmax loss as a weighted form:

$$R_1 = \sum\nolimits_i \log(\text{Softmax}(p_i)) \cdot s_i, \tag{3.12}$$

where p_i is the logit of candidate A_i, and s_i is the corresponding relevance score.

Binary Sigmoid Loss (R_2). This loss is close to the binary cross entropy loss:

$$R_2 = \sum\nolimits_i \left[\log(\sigma(p_i)) \cdot s_i + \log(\sigma(1-p_i)) \cdot (1-s_i)\right], \tag{3.13}$$

where σ is the Sigmoid function, p_i is the logit of candidate A_i, and s_i is the corresponding relevance score.

Generalized Ranking Loss (R_3). Note that the answer generation process can be viewed as a ranking problem. Therefore, we derive a ranking loss:

$$R_3 = \sum\nolimits_i \log \frac{\exp(p_i)}{\exp(p_i) + \sum_{j \in G} \exp(p_j)} \cdot s_i, \tag{3.14}$$

where p_i is the logit of candidate A_i, and G is a group of candidates that has a lower relevance score than A_i. s_i is equal to 1 when the corresponding relevance score is greater than 0 and s_i is equal to 0 when the corresponding relevance score is equal to 0. Note that this function is re-organized from ListNet [126].

Note that our loss functions are derived from the Eq (3.9), not just the regression of dense annotation. The comparison experiments will be given in Section 3.6.4.

3) Hidden Dictionary Learning. We find that the Eq (3.9) can be written as:

$$\sum\nolimits_u P_u(A) P(u|H) = \mathbb{E}_{[u|H]} \left[P_u(A) \right]. \tag{3.15}$$

Although we cannot determine the exact meaning of U, we try to use a vector representation $\boldsymbol{z}$ to approximate an expression of U. We can approximate $\mathbb{E}_{[U|H]} \left[P_u(A) \right]$ as NWGM $\left[P_u(A) \right]$ [124, 127] (*i.e.*, normalized weighted geometric mean), and this term can be further calculated by creating a dictionary D_z of $\boldsymbol{z}$:

$$\mathbb{E}_{[u|H]} \left[P_u(A) \right] \approx \text{Softmax}\{\boldsymbol{g}_z(\mathbb{E}_z \left[\boldsymbol{Z} \right])\}, \tag{3.16}$$

where $\boldsymbol{g}_z$ is a fully connected layer, $\boldsymbol{Z}$ represents a variable and its value $\boldsymbol{z}$ is selected from directory D_z. After deriving the last term, we can use D_z to calculate $\mathbb{E}_Z \left[\boldsymbol{Z} \right]$ shown in Figure 3.4(e) to approximate Eq (3.9). Noting that although when we train the dictionary, we still need to use answer score sampling, the hidden dictionary learning is a more proper way to approximate the unobserved confounder because it explores the whole space of U rather than the second attempt which only uses some samples of U.

3.6 Experiments

3.6.1 Experimental Setup

Dataset. Our principles are evaluated on the recently released real-world dataset VisDial v1.0[3]. Specifically, the training set of VisDial v1.0 contains 123K images from the COCO dataset [128] with 10 rounds of dialog for each image, a total of about 1.2M dialog pairs. The validation and test sets were collected from Flickr, with 2K and 8K COCO-like images respectively. The test set is further split into test-std and test-challenge splits, both with the number of 4K images that are hosted on the blind online evaluation server. Each image in training and validation sets has a 10-round dialog, while in the test set the number of the dialog is flexible. Every dialog in the VisDial dataset is given with 100 answer candidates. We evaluated our results on the validation and test-std set.

Metric. Normalized Discounted Cumulative Gain (NDCG) is appointed by the official and accepted by the community to evaluate the VisDial models. Note that NDCG assigns each answer candidate a relevance score based on whether the answer is suitable for the question (0 denotes the answer and higher relevance denotes the answer is more proper for the question), *e.g.*, there are 100 candidates for each question and NDCG assign relevance score for each of them and totally 100 relevance scores are given for this question. In the testing, NDCG requires the model to treat any suitable answer (relevance score bigger than 0) as correct, which is different from the training in that only one answer is given in the dialog to train the model, which essentially makes VisDias as an Out-of-Distribution Generalization task.

3.6.2 Model Zoo

We report the performance of the following baseline VisDial models, including LF [24], HCIAE [116], CoAtt [67] and RvA [25]:

LF [24]. The naive base model has no attention modules. We expand the model by adding some very basic attention operations, including question-based history attention and question-history-based visual attention refinement.

[3]Suggest by the official [112], results should be reported on v1.0 instead of v0.9

HCIAE [116]. The model consists of question-based history attention and question-history-based visual attention.

CoAtt [67]. The model consists of question-based visual attention, image-question-based history attention, image-history-based question attention, and the final question-history-based visual attention.

RvA [25]. The model consists of question-based visual attention and history-based visual attention refinement.

3.6.3 Implementation Details

Pre-processing. As for language pre-processing, we followed the process described by [24]. First, we lowercased all letters in the sentence, converted numbers to words and removed abbreviations. Afterward, we used the Python NLTK toolkit to tokenize the sentences into word lists, followed by padding or truncating the captions, questions, and answers to the length of 40, 20 and 20, respectively. And we built a token vocabulary of size 11,322 consisting of 11,319 words that occur at least five times in train v1.0 and three instruction tokens. We initialized all word embeddings by loading GloVe [129], which were shared between the encoder and decoder, and we applied 2-layers LSTM to encode word embedding and set their hidden dimension to 512. As for the visual features, we used the bottom-up-attention features [57] given by the official [112].

Implementation of Principles. For P1, we removed history features from the final fusion vector representation of all models, while leaving the rest unchanged. For HCIAE [116] and CoAtt [67], we also masked the history guidance to the image. For P2, we trained our models using the preference scores computed from the question types or given by the official (*i.e.*, dense annotation in train v1.0). Specifically, for "question type", we first defined 55 types and marked answers that occurred over 5 times as preferred answers, and then used the preference scores to train our model by R_2 loss. "Answer score sampling" was directly used to train our pre-trained model by the proposed loss function. For "dictionary", we set up a memory of dimension 100×512 to implement D_z, then trained it via dense annotations by R_3 loss. Note that other implementations different from ours are also acceptable as long as following P1 and P2.

Model	baseline	QT	S				D
			R_0	R_1	R_2	R_3	
LF [24]	57.21	58.97	67.82	71.27	72.04	72.36	72.65
LF +P1	61.88	62.87	69.47	72.16	72.85	73.42	**73.63**

TABLE 3.1: Performance (NDCG%) comparison for the experiments of applying our principles on the VisDial v1.0 validation set. LF is the enhanced version as we mentioned. QT, S and D denote question type, answer score sampling, and hidden dictionary learning, respectively. R_0, R_1, R_2, R_3 denote regressive loss (baseline applying relevance score), weighted softmax loss, binary sigmoid loss, and generalized ranking loss, respectively.

Training. We trained the model with P1 using the Softmax cross-entropy loss with Adam [130], where the learning rate of 4×10^{-3}, decaying at epoch 5, 7, 9 with a decay rate of 0.4. We trained the model for a total of 15 epochs. Dropout [127] was also applied, with a rate of 0.4 for the RNN and 0.25 for the fully connected layer. Other settings were set by default.

3.6.4 Quantitative Results

Table 3.1 shows the results of different implementations in P2, *i.e.*, question type, answer score sampling, and hidden dictionary learning. Overall, all implementations improve the performance of the base models. Specifically, the attempts of P2 can further boost performance by 11.75% at most by hidden dictionary learning. More specifically, our designed loss functions based on Eq. (3.8) outperform the regressive score (*i.e.*, R_0) which is a naive Euclidean distance loss, which can be seen as the baseline of using the additional relevance score, and we also find that our proposed generalized ranking loss (*i.e.*, R_3) is the best because it satisfies the ranking property of VisDial.

To demonstrate that our principles are model-agnostic, Table 3.2 shows the results of our experiments about applying our principles on four different models (*i.e.*, LF [24], HCIAE [116], CoAtt [67] and RvA [25]). In general, our two principles can improve all the models in any ablative condition. We also find that the effectiveness of P1 and P2 are additive, *i.e.*, their combination performs the best. Note that the enhanced LF model is very simple, with no complex attention strategies. However, this simple architecture with our principles still does not hinder it to achieve the best performance.

Model	LF [24]	HCIAE [116]	CoAtt [67]	RvA [25]
baseline	57.21	56.98	56.46	56.74
+P1	61.88	60.12	60.27	61.02
+P2	72.65	71.50	71.41	71.44
+P1+P2	**73.63**	71.99	71.87	72.88

TABLE 3.2: Performance(NDCG%) of ablative studies on different models on VisDial v1.0 validation set. P2 indicates the most effective one (*i.e.*, hidden dictionary learning) shown in Table 3.1. Note that only applying P2 is implemented by the attempts in Section 3.5 with the history shortcut.

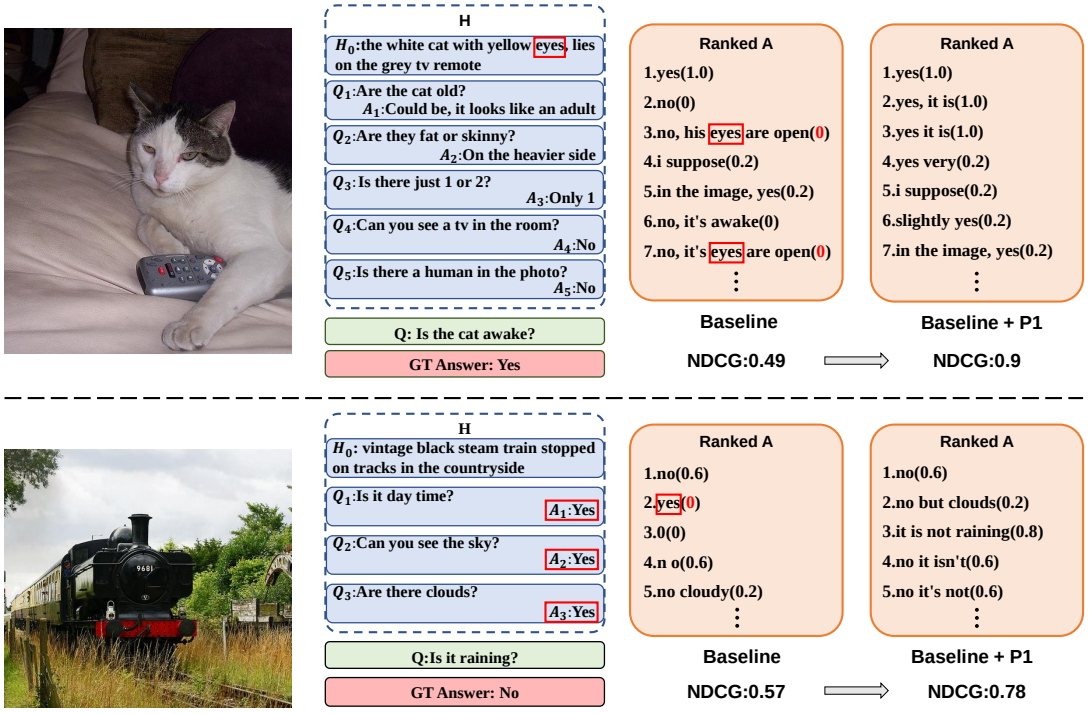


FIGURE 3.5: Qualitative results for baseline and baseline with P1 on the VisDial v1.0 validation set. Numbers in brackets in ranked A indicate relevance scores. The red boxes indicate that the baseline model replicates the words from the dialog history, even if they are literally nonsense for answering the current question. The bottom example shows that although the baseline can correctly select the ground truth answer, it is influenced by the history bias, and thus it ranks "yes" at a high place, which degrades its performance (NDCG). As for the baseline with P1, it does not make such unreasonable choices.

3.6.5 Qualitative Analysis

The qualitative results illustrated in Figure 3.5 and Figure 3.6 show the following advantages of our principles.

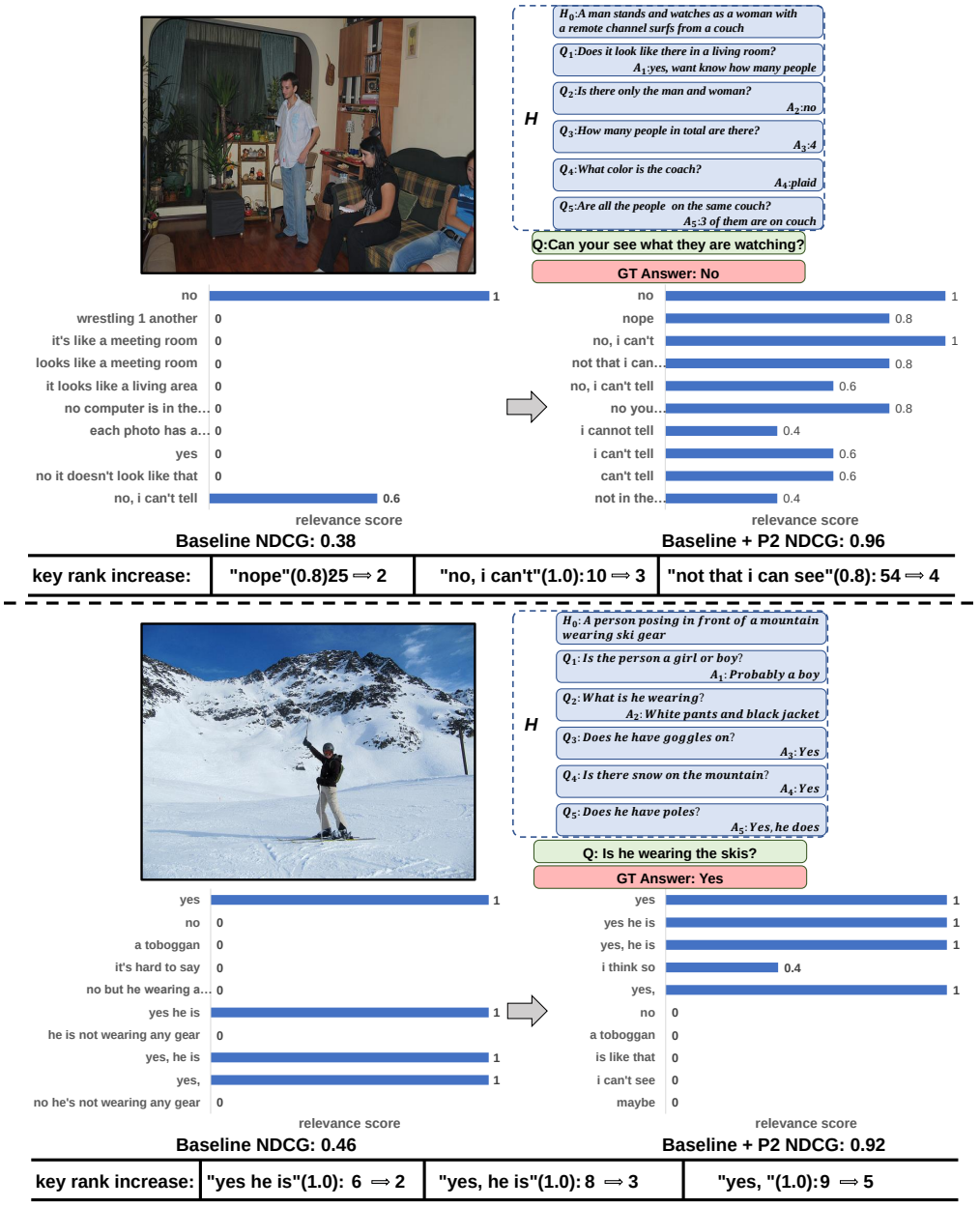


FIGURE 3.6: Qualitative examples of the ranked candidates of baseline and baseline with P2. We also give some key rank changes for boosting NDCG performance by implementing P2. These examples are taken from the validation set of VisDial v1.0.

Removing History Bias. After applying P1, the harmful bias learned from history is mitigated, especially the answer-length bias shown in Figure 3.1(a) and word-match bias shown in Figure 3.5. After applying P1, the average length of top-1 answers (*i.e.*, the blue line in Figure 3.1(a)) is no longer related to the average length of history answers and becomes closer to the average length of NDCG ground truth answers (*i.e.*, green dash line). As for the word-match bias in Figure 3.5,

we can observe that the word "eyes" in history is literally unrelated to the current question. However, at the top of the ranked answer list of the baseline model, the word "eyes" can be found in some undesirable candidates (*i.e.*, with low relevance score). In general, the baseline model prefers to match words in the history and rank matching candidates high due to the direct path from the history to the answer. If we count the number of matches from the baseline with meaningful words in the top 10 candidates obtained from P1 and the baseline (*e.g.*, the word "eyes"), we find that P1 can match about 10% fewer words from the history ($\sim$ 4800 times compared to $\sim$ 5200 times).

The bottom example shown in Figure 3.5 also illustrates a type of word match. In the ranked list of the baseline model, "yes" is ranked very high, and "yes" exists in history many times. By analyzing the validation results, we found that if "yes" or "no" exists in dialog history, the baseline model will give them an above-average ranking due to word matching. With the application of P1, this phenomenon will no longer occur.

More Reasonable Ranking. Figure 3.6 shows that the baseline model is only concerned with ground truth answer like "no" or "yes" and not with the rank of other answers with similar semantics like "nope" or "yes, he is". This is inconsistent with human intuition, as we assume that candidates with similar semantics are still the correct answers. This also results in the baseline model performing poorly under the NDCG metric. Compared to the model with P2, in the bottom example, it almost ranks all the appropriate answers such as "yes, he is", "yes he is" and "I think so" at the top along with the ground truth answer "yes", which greatly improves the NDCG performance.

3.6.6 Visual Dialog Challenge

Finally, we used the online blind test server to demonstrate the effectiveness of our principles on the VisDial v1.0 test-std set. As shown in Table 3.3, the top part contains the results of the baseline models implementing our principles, where P2 denotes the most effective one (*i.e.*, hidden dictionary learning). The bottom part is the 2019 Visual Dialog Challenge leaderboard [10]. We used the ensemble of the enhanced LF [24] to beat our best performance (MReaL-BDAI) in the 2019 Visual Dialog Challenge, which also used other implementations of P1 and P2.

	Model	NDCG(%)
Ours	P1+P2 (More Ensemble)	**74.91**
	LF+P1+P2 (Ensemble)	74.19
	LF+P1+P2 (single)	71.60
	RvA+P1+P2 (single)	71.28
	CoAtt+P1+P2 (single)	69.81
	HCIAE+P1+P2 (single)	69.66
Leaderboard	MReaL-BDAI*	74.02
	ReDAN+ (Ensemble) [63]	64.47
	square*	60.16
	VIC-SNU [66]*	57.59
	UET-VNU*	57.40
	idansc [64]*	57.13

TABLE 3.3: Our results and comparisons to the 2019 Second Visual Dialog Challenge Leaderboard results on the test-std set of VisDial v1.0. Results are reported by the test server, (*) is taken from [10].

Promisingly, by applying our principles, we can promote all the single baseline models to the top ranks on the leaderboard. Note that, based on our principles, we were awarded the winner of the 2019 Visual Dialog Challenge and the runner-up of the 2020 Visual Dialog Challenge.

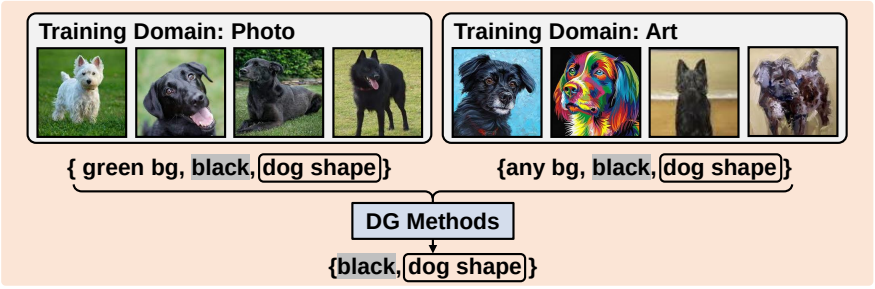
Chapter 4

OOD Generalization with Multiple Domains[1]

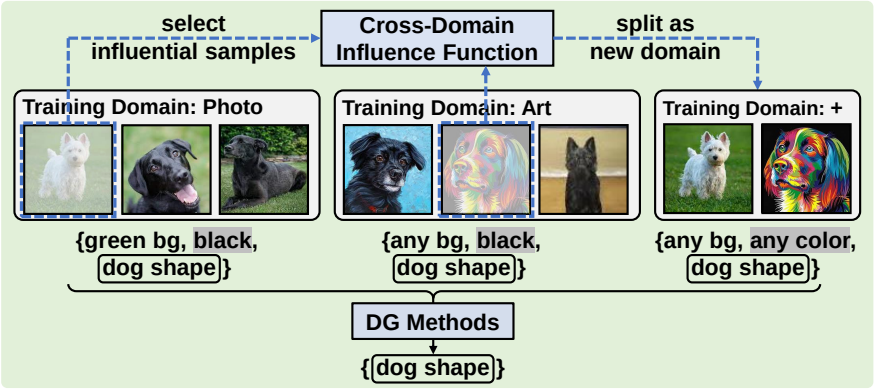
4.1 Case Study: Domain Generalization

Deep models are good at fitting training data but are not suitable for generalizing to unseen domains [12, 14–16], which have different data distribution of the training. For example, given `Photo` domain, where most dogs are black, and train a model, the model will learn color features to classify dogs regardless of the dog's features (*i.e.*, color is learned as bias in this domain), and the result is it has the less discriminative ability when the color is no longer needed, *e.g.*, tested in `Sketch` domain. In practice, models are usually tested in various domains, and thus we are interested in the Domain Generalization (DG) task: given multiple domains and train a model to realize *invariance*, which means the model can generalize to any unseen domain [131, 132]. Note that, in this chapter, all methods need the domain labels, which are additionally provided by the dataset beside the classification annotation, and this is different from the setting we will discuss in the Chapter 5, which can be treated as (Implicit) Debiasing case, where no additional annotation is given in the training.

[1]The main content in this chapter is submitted to *CVPR 2023* as **Jiaxin Qi***, Zike Wu*, Qianru Sun, Xian-Sheng Hua, and Hanwang Zhang. DOMAIN+: Splitting a New Influential Domain for Domain Generalization. *Under Review*. The superscript $*$ indicates equal contributions.



(a) DG methods learn spurious invariance



(b) DG methods with our DOMAIN+ learn invariance

FIGURE 4.1: Illustration of the spurious invariance (grey shade), inducing domain bias, learned by traditional DG methods and removed by our DOMAIN+. "bg" denotes background and bordered words, such as "dog shape", denote domain-invariant features.

To this end, all DG methods aim to keep the domain-invariant features (*i.e.*, causal features [17, 105]) by discarding the domain-specific ones [4, 32, 77], which induce the domain bias. As shown in Figure 4.1(a), {`black, dog shape`} are the invariant features obtained by DG methods as they are indeed discriminative for most training samples in both `Photo` and `Art`. Embarrassingly, the community recently finds out that the most naive Empirical Risk Minimization (ERM) objective, which simply merges the samples of all the training domains into one training set without applying any domain-invariant strategy, shows comparable or even better performance compared to DG methods [12] with complex invariance-learning design. The reason is that although ERM is widely known to be easily biased by the spurious correlations in the training [4, 15, 94, 96, 97], *e.g.*, although most dogs are black in one domain, only if the training samples across other domains are diverse, *e.g.*, dogs with different colors in `Photo` and `Art`, ERM can still remove domain-specific features (*i.e.*, remove the (single) domain bias, where the bias exists in one domain but not in the whole combined dataset), *e.g.*, it learns {`black, dog shape`} as

FIGURE 4.2: Visualizations of the samples ranked by cross-domain influence (top) and IRM loss (bottom) from low to high. Red borders denote the selected rare samples by DOMAIN+.

well as DG methods.

This implies that the traditional DG methods still lack a self-diagnostic mechanism to remove another type of domain bias called *Spurious Invariance*, which is invariant across all training domains but variant in the testing domain. We call bias as "spurious invariance" because it cannot be overturned by using their invariant strategy with the cross-domain validation only on the training domains [4, 73, 77], which has the same behavior, just like the underlying true invariance that is also shared by all training domains, removing which will definitely increase the training loss. In Figure 4.1(a), {`black`} is a spurious invariance because it is not discriminative in unseen domains without color such as `Sketch`. The reason is that the {`black`} samples prevail in all training domains over {`other color`} samples, *e.g.*, the white in `Photo` and the colorful in `Art`. Thus, these rare samples make minor contributions to counter that {`black`} is not the true invariance, because removing it will diminish the model fitting for the majority "black dog" training samples.

Someone may propose a straightforward solution which is to up-weight those rare samples. This sounds appealing but it is difficult to apply in practice due to the following two challenges. First, it is hard to identify the sample-wise "rarity" as the notion of "invariance loss", which is defined on the dataset level to evaluate how consistently the model (or feature) behaves on this dataset/domain distribution [4, 76, 77]. So, popular hard sample mining methods only identify those with high training loss [82, 109, 133], which only quantifies how well the model fits a specific sample but cannot reflect the level of invariance [2, 81]. Second, even if we can accurately zoom into those rare samples, the re-sampling of them will introduce not only the desired features (*e.g.*, {`white`} in `Photo`), but also other associated ones

(*e.g.*, {`green bg`}), which may cause new bias misleading the entire training [6, 134], and we will discuss this problem in detail in Section 5.2.

To address the first challenge, a possible way to judge if a sample has no spurious invariance is to ask the self-diagnostic counterfactual question: If we have removed the sample, how would it affect the invariance re-trained on the new training data excluding it? In particular, we define the answer as a sample-wise real value called *Cross-domain Influence*: as the samples without spurious invariance are rare, if we remove one of them in one domain, the domain's spurious invariance will become more dominant, *e.g.*, the percentage of black dogs in `Photo` is higher, then the spurious invariance will be more easily achieved in the domain after training. So, such "purer" spuriousness helps other domains achieve the spurious invariance faster too—decreasing their invariance losses; in contrast, if we remove one of the majority samples with spurious invariance, it won't significantly decrease the invariance loss as the spurious invariance is still dominant. We provide a formal justification in Section 4.3.2 to show that the cross-domain influence is a good measure for rare samples.

However, the above "leaving one sample out and re-training" makes the cross-domain influence estimation cost a lot. Thanks to the recent advances in approximating the sample influence without re-training by influence function [88, 99], we can implement our cross-domain influence by "differentiating" a sample in one domain, *i.e.*, up-weighting the sample with an infinitesimal amount, and then estimating the mean of the invariance loss changes in each other domains by a closed-form expression (The formula can be found in Section 4.3). As shown in Figure 4.2, ranking the samples by their influence indeed tells us more about the spurious invariance than the conventional sample "hardness". For example, our high influence identifies rare dogs that are {`non-standing, abnormal action, colorful`}, which do not suggest high training loss necessarily.

Finally, to address the **second challenge**, after identifying the rare samples by their influence score, instead of re-sampling, we treat them as a new domain by splitting them from the original ones. Then, we can use any off-the-shelf DG methods on the old domains (without the split samples) plus the new one, and hence we dub our method DOMAIN+. As illustrated in Figure 4.1(b), DOMAIN+ can help any DG method to remove the bias caused by spurious invariance and achieve the true invariance {`dog shape`}, which is the only invariance across the newly

split domains. In Section 4.4, we use three classic open-sourced SOTAs: IRM [4], CORAL [32], and Fish [77], as our baselines on four popular DG datasets: PACS, VLCS, OfficeHome, and TerraIncognita. Specifically, we follow DomainBed [12]—a stringent and reproducible DG benchmark—to conduct all the experiments. The results show that we can consistently improve all the baselines, demonstrating that our Domain+ helps DG methods achieve a better invariance.

4.2 Preliminaries: ERM and IRM in DG

Given training data $\mathcal{D}$ consisting of K domains $\mathcal{D} = \{\mathcal{D}_1, \mathcal{D}_2, ..., \mathcal{D}_K\}$, where $\mathcal{D}_k = \{(x_i^k, y_i^k)\}_{i=1}^{n_k}$, x_i^k is a sample in domain k, y_i^k is its one-hot label, and n_k is the number of samples in $\mathcal{D}_k$. Domain Generalization (DG) aims to train a model f on $\mathcal{D}$ to predict the labels of testing samples in any unseen domains $\mathcal{D}_u$. The crux of learning f is to capture the domain-invariant (causal) features, which are invariantly discriminative in any domain, by discarding all the domain-specific features that are only discriminative in training but not testing.

Empirical Risk Minimization (ERM). It simply merges the samples of all the training domains as a whole without domain index, *i.e.*, $\mathcal{D} = \bigcup_{k=1}^{K} \mathcal{D}_k = \{(x_i, y_i)\}_{i=1}^{n}$, where $n = \sum_{k=1}^{K} n_k$. ERM learns f on $\mathcal{D}$ by minimizing the softmax cross-entropy (CE) loss:

$$\mathcal{L}_{\text{ERM}}(f, \mathcal{D}) = \frac{1}{n} \sum_{i=1}^{n} \text{CE}(y_i, f(x_i)), \tag{4.1}$$

where $f(x_i)$ is the softmax prediction of x_i. ERM can remove some domain-specific features and shows competitive performance as we mentioned in Section 4.1. The reason is that some domain-specific features are no longer dominant in the combined set. For example, in Figure 4.1(a), although `green bg` is a domain-specific feature in `Photo`, it is less dominant in `Photo` and `Art` combined. However, ERM cannot remove dataset-specific features when most domains contain similar features, which still causes bias in unseen testing domains.

Invariant Risk Minimization (IRM). Domain Generalization (DG) methods aim to discard the domain-specific features by additionally minimizing a penalty across all training domains, *i.e.*, the *invariance loss* $\mathcal{L}(f, \mathcal{D}_k)$:

$$\mathcal{L}_{\mathrm{DG}}(f) = \frac{1}{K}\sum_{k=1}^{K}[\mathcal{L}_{\mathrm{ERM}}(f, \mathcal{D}_k) + \lambda \cdot \mathcal{L}(f, \mathcal{D}_k)], \tag{4.2}$$

where $\lambda > 0$ is a trade-off hyper-parameter. For example, Invariant Risk Minimization (IRM) [4], a classic DG method, implements the invariance loss as:

$$\mathcal{L}(f, \mathcal{D}_k) = \sum_{i=1}^{n_k} \|\nabla_{\theta|\theta=1}\, \mathrm{CE}(y_i^k, f(x_i^k) \cdot \theta)\|^2, \tag{4.3}$$

where θ is a "dummy" classifier, whose gradient is not used to update itself but to calculate the penalty. Invariance loss encourages the model to be equally optimal in different training domains, by penalizing the learning of different domain-specific features. Note that the invariance loss is a dataset-level loss, which should be measured on the whole domain/dataset and its value for a single sample is meaningless.

However, DG methods in the form of Eq. (4.2) cannot remove the domain-specific features shared by all training domains, leaving the *spurious invariance*, which is invariant across training domains but variant to the testing domains. The reason is that $\mathcal{L}(f, \mathcal{D}_k)$ is essentially a pooling of domain samples, and in this way, the contribution of some rare samples without the spurious invariance is thus suppressed.

4.3 Our Algorithm: Domain+

To help DG methods overcome the bias induced by the spurious invariance, we propose DOMAIN+: 1) find the rare samples without spurious invariance by the proposed cross-domain influence, 2) split them from their original domains as a new domain, and then train DG methods on the original domains, without the split samples, plus the new one. DOMAIN+ algorithm is summarized in Algorithm 1.

4.3.1 Implementations

Cross-Domain Influence. As we discussed in Section 4.1, the sample "rarity" cannot be identified by the dataset-level loss such as Eq. (4.3). To this end, we introduce a sample-level index called *cross-domain influence* for sample x from

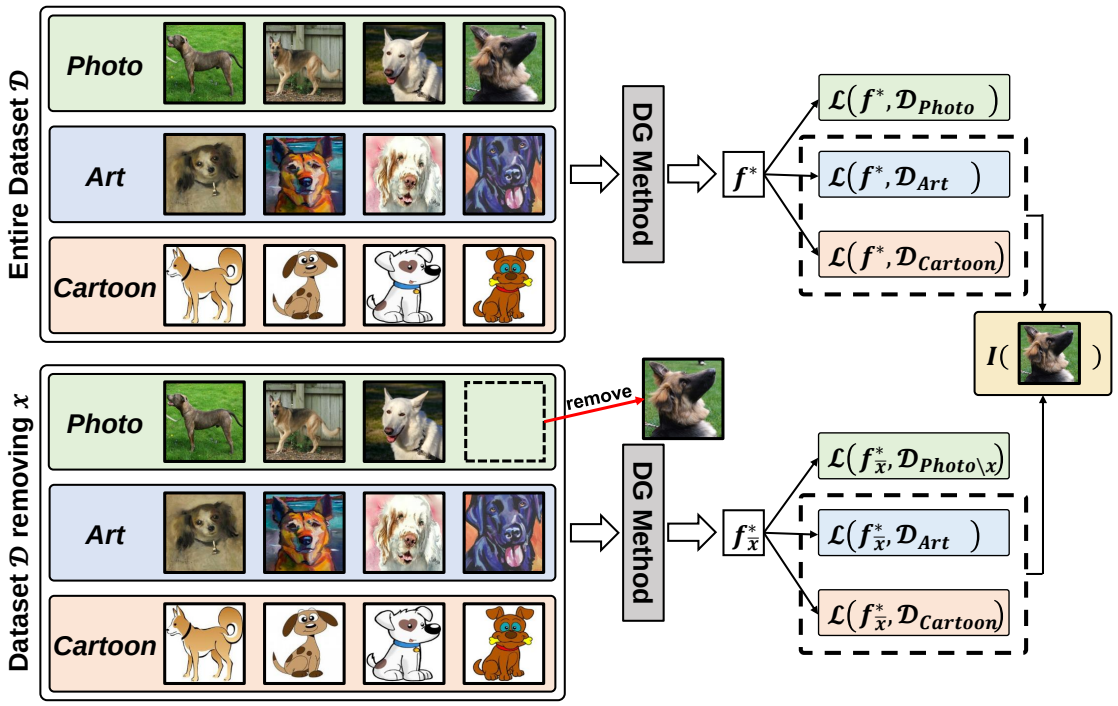


FIGURE 4.3: Illustration of the calculation of the cross-domain influence in Eq. (4.4), where the most right part denotes the difference between the two loss values (illustrated in dashed boxes).

domain k_x:

$$I(x) = \frac{1}{K-1} \sum_{k=1, k\neq k_x}^{K} \mathcal{L}_k(f^*) - \frac{1}{K-1} \sum_{k=1, k\neq k_x}^{K} \mathcal{L}_k(f^*_{\bar{x}}), \tag{4.4}$$

where $\mathcal{L}_k(f) := \mathcal{L}(f, D_k)$, f^* and $f^*_{\bar{x}}$ denote the optimal model trained on the entire dataset $\mathcal{D}$ and $\mathcal{D}\backslash\{x\}$, respectively. The term "cross-domain" means that the sample removal happens in its own domain but its counterfactual influence is calculated by the mean of the invariance changes across other domains. The calculation details are illustrated in Figure 4.3. However, Eq. (4.4) needs to re-train the model on the new dataset $\mathcal{D}\backslash\{x\}$ that is prohibitively expensive.

Thanks to the recent advances in approximating the sample influence without re-training [88, 98], we can implement $I(x)$ by "differentiating" a sample x from domain k_x to derive the gradients of the invariance loss of other domains, *i.e.*, by

only training once, we can effectively estimate the influence for each sample:

$$\begin{aligned} I(x) &= \frac{1}{K-1}\sum_{k=1,k\neq k_x}^{K} \left.\frac{\mathrm{d}\mathcal{L}_k(f_\epsilon^*)}{\mathrm{d}\epsilon}\right|_{\epsilon=0} \\ &= -\frac{1}{K-1}\sum_{k=1,k\neq k_x}^{K} \nabla\mathcal{L}_k(f^*)H_{f^*}^{-1}\nabla\mathcal{L}(f^*,x), \end{aligned} \tag{4.5}$$

where ϵ denotes an infinitesimal perturbation, $f_\epsilon^* := \arg\min_f \mathcal{L}_{\mathrm{DG}}(f) + \epsilon\mathcal{L}(f,x)$ is the optimal model after perturbing x, where $\mathcal{L}(f,x) = \|\nabla_{\theta|\theta=1}\,\mathrm{CE}(y, f(x)\cdot\theta)\|^2$ if we implement IRM as the invariance loss, the perturbation only happened on $\mathcal{L}(f,x)$ due to we are interested in the changes of the invariance loss but not the whole loss, and $H_{f^*} := \nabla^2\mathcal{L}_{\mathrm{DG}}(f^*)$ denotes the Hessian matrix, which derives from the influence function [88].

Note that the sample-wise gradient of invariance loss reflects the domain changes and it is meaningful for domain-level invariance, which is different from the sample-wise loss we have discussed. Our cross-domain influence function is calculated on other training domains, which is more reasonable than the original one based on the testing set.

Rare samples split into a new domain. After estimating the cross-domain influence of each sample x, we split the rare samples by $I(x) > \alpha$ from their original domain, and construct a new domain $\mathcal{D}^+$, where α is a threshold. Then, we train DG methods on $\{\mathcal{D}_k \backslash \mathcal{D}^+\}_{k=1}^K \cup \{\mathcal{D}^+\}$ to achieve the invariance. We'd like to highlight again that the influence is fundamentally different from the sample hardness in hard example mining [133, 135]. Besides the qualitative samples in Figure 4.2, we also show the feature distributions of all the samples of different classes in Figure 4.5. Interestingly, we can see the difference of $\mathcal{D}^+$ selected by influence and training loss: as the rare samples with large influence are usually confounded by the majority, they are more evenly distributed than the "hard" samples, which are merely considered as the eccentric points far from the mainstream.

4.3.2 Justifications

We first show that rare samples indeed have larger cross-domain influence by Theorem 4.3, and then demonstrate that splitting the rare samples into a new domain

Algorithm 1: DOMAIN+

Input : Dataset $\mathcal{D} = \{\mathcal{D}_k\}_{k=1}^K$, Threshold α
Output: New Dataset $\{\mathcal{D}_k \setminus \mathcal{D}^+\}_{k=1}^K \cup \{\mathcal{D}^+\}$
Train f on $\mathcal{D}$ by Eq. (4.2) and derive the optimal f^*;
Initialize: $\mathcal{D}^+ \leftarrow \varnothing$;
foreach $\mathcal{D}_k \in \mathcal{D}$ **do** // Enumerate domains
  Initialize: $\mathcal{D}_r \leftarrow \varnothing$; // Rare sample set
  foreach $x \in \mathcal{D}_k$ **do**
    Initialize: $I(x) \leftarrow 0$;
    foreach $\mathcal{D}_j \in \mathcal{D} \setminus \{\mathcal{D}_k\}$ **do**
      $I(x) \leftarrow I(x) - \nabla\mathcal{L}_j(f^*) H_{f^*}^{-1} \nabla\mathcal{L}(f^*, x)$; // Eq. (4.5)
    $I(x) \leftarrow I(x)/(K-1)$
    if $I(x) > \alpha$ **then**
      $\mathcal{D}_r \leftarrow \mathcal{D}_r \cup \{x\}$; // x is rare
  $\mathcal{D}^+ \leftarrow \mathcal{D}^+ \cup \mathcal{D}_r$; // Update DOMAIN+
Apply any DG methods on $\{\mathcal{D}_k \setminus \mathcal{D}^+\}_{k=1}^K \cup \{\mathcal{D}^+\}$.

helps DG methods achieve better invariance by Theorem 4.4. We use $\mathcal{D}_d = \{x_d\}$ to denote the set of dominant samples from all training domains and $\mathcal{D}_r = \{x_r\}$ to denote rare samples set.

First, we prove $I(x_d) = 0$ by Lemma 4.1, $I(x_r) > 0$ by Lemma 4.2, and finally derive $I(x_r) > I(x_d)$ in Theorem 4.3.

Lemma 4.1. *Let f^* be the optimal model trained by DG methods, which learns the spurious invariance,* i.e., *achieves the local minimum of invariance loss on $\mathcal{D}_d$. Then, for all dominant sample $x_d \in \mathcal{D}_d$, we have $I(x_d) = 0$.*

Proof. As f^* achieves local minimum of invariance loss on $\mathcal{D}_d$, then for all $x_d \in \mathcal{D}_d$, we have $\nabla\mathcal{L}(f^*, x_d) = 0$. Therefore, we derive the following equation according to Eq. (4.5):

$$\begin{aligned} I(x_d) &= \frac{1}{K-1} \sum_{k=1, k\neq k_x}^{K} \nabla\mathcal{L}_k(f^*) H_{f^*}^{-1} \nabla\mathcal{L}(f^*, x_d) \\ &= \frac{1}{K-1} \sum_{k=1, k\neq k_x}^{K} \nabla\mathcal{L}_k(f^*) H_{f^*}^{-1} \cdot 0 = 0, \end{aligned}$$

where k_x is the domain index of x_d. □

Lemma 4.2. *Let f^* be the optimal model defined in Lemma 4.1. Then, for all rare samples $x_r \in \mathcal{D}_r$, we have $I(x_r) > 0$.*

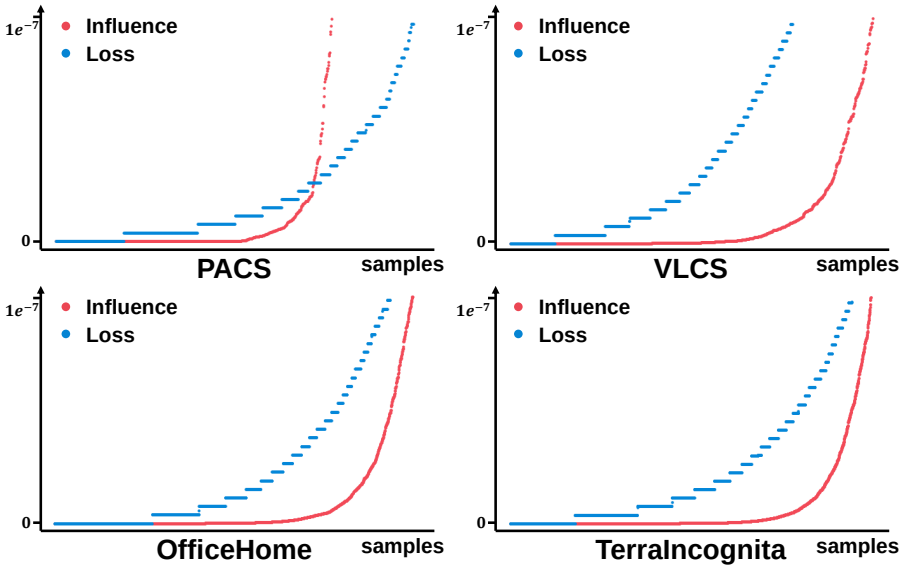


FIGURE 4.4: Visualization of sorted cross-domain influence (red) and training loss (blue) of training samples using IRM. We train the model on the default three training domains for each dataset. Each dot denotes a sample and its influence/loss value.

Proof. Because f^* is the local minimum of the invariance loss on $\mathcal{D}_d$, if we up-weight a rare sample $x_r \notin \mathcal{D}_d$ by ϵ and retrain f, the new optimal model f^*_ϵ will deviate from f^* with respect to the invariance loss.

Therefore, for each domain k, we have

$$\mathcal{L}_k(f^*_\epsilon) > \mathcal{L}_k(f^*).$$

When ϵ is approaching 0, according to Eq (4.5) we have

$$\begin{aligned} I(x_r) &= \frac{1}{K-1} \sum_{k=1,k\neq k_x}^{K} \frac{\mathrm{d}\mathcal{L}_k(f^*_\epsilon)}{\mathrm{d}\epsilon}\bigg|_{\epsilon=0} \\ &= \frac{1}{K-1} \sum_{k=1,k\neq k_x}^{K} \lim_{\epsilon\to 0} \frac{\mathcal{L}_k(f^*_\epsilon) - \mathcal{L}_k(f^*)}{\epsilon} > 0, \end{aligned}$$

where k_x is the domain index of x_r.

□

Theorem 4.3. *For all rare sample $x_r \in \mathcal{D}_r$ and all dominant sample $x_d \in \mathcal{D}_d$, the cross-domain influence of x_r is larger than x_d,* i.e., *$I(x_r) > I(x_d) = 0$.*

The proof can be directly obtained based on Lemma 4.1 and Lemma 4.2. As shown in Figure 4.4, Theorem 4.3 works well in practice on four datasets—most of the dominant samples' influence scores are indeed close to 0 and the turning curve from dominant to rare is much sharper than that of using training loss.

Then, we split rare samples into a new domain according to Theorem 4.3. In practice, we select $\mathcal{D}^+ = \{x | I(x) > \alpha\}$, where α is a threshold slightly greater than 0 that tolerates the estimation error of $I(x)$.

Theorem 4.4. *$\mathcal{D}^+$ as a new training domain can reduce the degree of freedom in the invariant solution space* w.r.t.*the learned invariant features.*

Here, the degree of freedom (DOF) indicates the rank of learned invariant features. The reduction of DOF is equivalent to the removal of spurious invariant features, which means the model has a better generalization ability [4]. Hence, Theorem 4.4 says that DomainBed+ will help DG methods achieve better invariance.

4.4 Experiments

4.4.1 Experimental Setups

Dataset. Following DomainBed, we demonstrated our DomainBed+ on 4 popular multi-domain image classification datasets. 1) PACS [11] contains 9,991 images of 7 classes from four domains, including `art`, `cartoons`, `photos`, and `sketches`. 2) VLCS [136] contains 10,729 photographs of 5 classes from four domains, including `Caltech101`, `LabelMe`, `SUN09`, and `VOC2007`. 3) Office-Home [137] contains 15,588 images of 65 classes from four domains, including `art`, `clipart`, `product` and `real`. 4) Terra Incognita [138] contains 24,788 photos of wild animals from 10 classes, taken at four different locations, including `L100`, `L38`, `L43` and `L46`.

DomainBed Benchmark [12]**.** It is a rigorous and reproducible domain generalization testbed that provides a consistent implementation of each SOTA method for fair comparisons. We follow a training-domain validation set, which is suggested in DomainBed, for model selection. Specifically, we split each training domain into training and validation subsets, accounting for 80% and 20%, respectively. We chose the model with the highest average accuracy on the validation sets. For fair

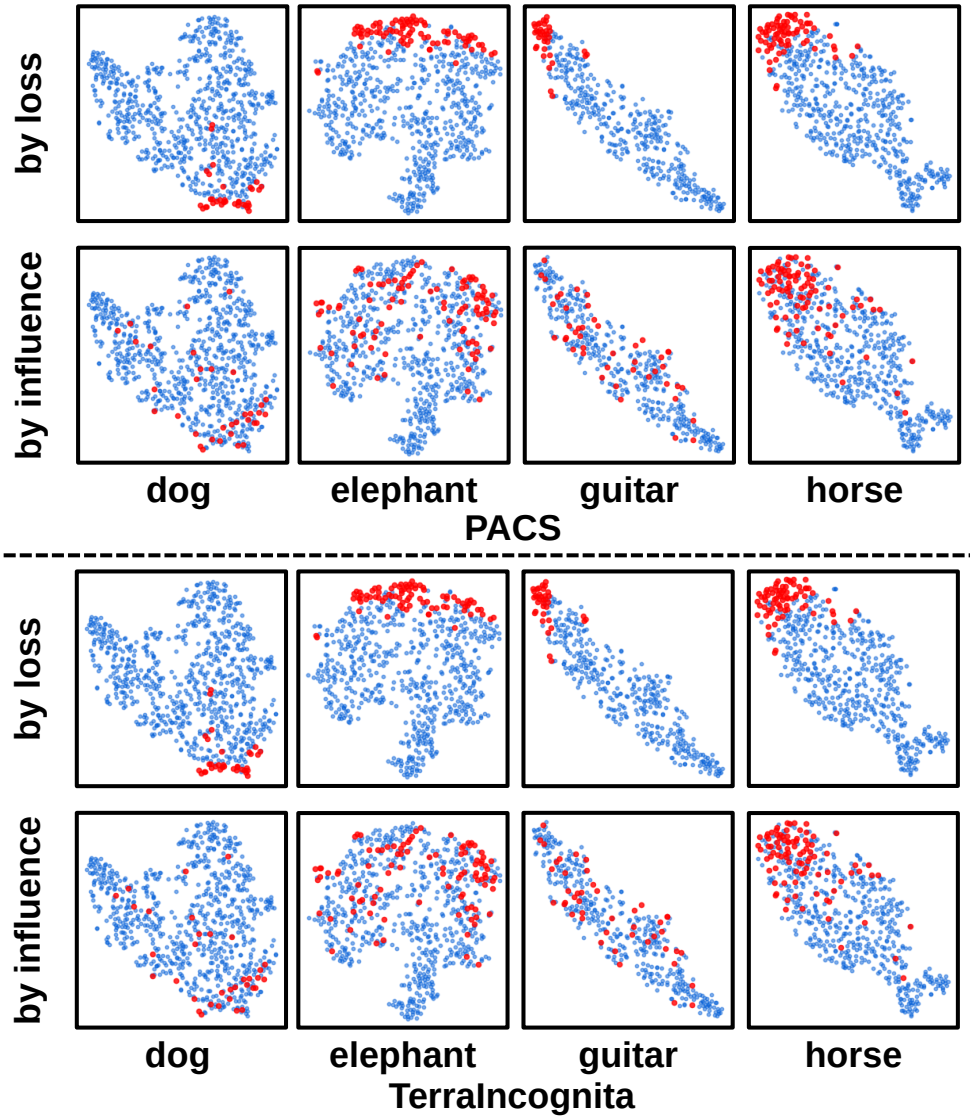


FIGURE 4.5: t-SNE [1] visualization of the training sample features extracted by IRM model. We trained the model on the default three domains on each dataset. Red dots are the selected rare samples by training loss and cross-domain influence.

comparisons, we followed the same settings in Fish [77] that report the average over five random trials.

Baselines. We chose 3 popular DG SOTAs: IRM [4], CORAL [32], and Fish [77], and applied our DOMAIN+ to them, where we later named IRM+, CORAL+, and Fish+, respectively. We compared their performances with other SOTAs based on the implementation of DOMAINBED, including ERM [139], DRO [74], Mixup [103], MLDG [70], MMD [76], RSC [69], ANDMask [78] and SagNet [140].

Algorithm	PACS					VLCS				
	A	**C**	**P**	**S**	**Avg**	**C**	**L**	**S**	**V**	**Avg**
ERM [139]	**86.7**±1.3	79.6±2.7	95.8±0.6	79.2±2.6	85.3±1.3	97.1±1.0	**65.7**±1.5	69.7±2.9	74.3±3.6	76.7±1.2
DRO [74]	83.6±2.2	79.7±2.3	96.5±0.4	78.9±2.3	84.7±1.3	96.9±1.2	63.3±1.1	70.0±2.5	72.9±2.7	75.8±1.4
Mixup [103]	85.3±1.1	80.5±1.2	96.9±0.3	75.9±2.9	84.6±1.1	97.6±0.7	63.2±1.5	70.6±1.6	74.9±1.4	76.6±0.9
MLDG [70]	83.0±4.9	76.2±1.8	95.8±1.1	74.5±2.0	82.4±1.4	97.2±0.9	63.2±2.2	70.1±2.1	72.5±1.6	75.7±1.1
MMD [76]	83.4±2.1	79.4±3.7	95.7±0.7	74.0±7.0	83.1±2.3	97.4±0.9	62.9±2.0	69.9±1.8	74.8±3.1	76.2±1.5
RSC [69]	80.6±2.9	77.5±3.4	95.1±0.6	76.9±2.7	82.5±1.4	93.7±1.8	64.2±1.8	67.8±1.4	71.1±3.5	74.2±1.0
ANDMask [78]	84.3±3.1	77.6±1.9	96.3±0.7	72.7±4.4	82.7±2.3	96.7±1.4	63.9±2.1	67.1±3.3	70.4±3.1	74.5±1.7
SagNet [140]	83.2±0.6	81.1±1.2	95.5±1.2	77.9±2.2	84.4±0.8	96.1±1.3	63.3±2.3	72.3±3.4	73.7±2.7	76.3±0.9
IRM [4]	85.7±2.1	79.8±1.6	95.8±0.4	78.0±1.7	84.8±0.4	94.7±3.1	64.7±1.5	70.2±1.3	73.8±3.7	75.9±0.8
IRM+	85.9±2.8	81.1±0.3	96.7±0.7	78.7±1.4	85.6±0.7	97.2±0.8	65.5±1.7	71.3±2.0	75.9±1.2	77.5±0.8
CORAL [32]	84.2±2.4	78.8±3.1	96.6±0.6	77.5±1.3	84.3±0.8	97.1±0.5	65.5±1.2	70.3±2.5	**76.8**±2.2	77.4±0.8
CORAL+	86.5±2.0	**81.3**±2.3	**97.0**±0.6	**80.8**±0.8	**86.4**±0.7	**98.3**±0.6	65.3±1.6	71.6±1.9	76.5±2.2	77.9±0.7
Fish [77]	85.3±1.8	79.0±1.3	95.9±1.0	78.3±2.8	84.6±1.2	97.5±1.1	64.2±1.6	71.2±0.7	75.4±1.4	77.1±0.6
Fish+	86.1±1.8	81.1±2.0	96.9±0.8	78.7±3.4	85.7±1.0	98.0±0.9	65.3±1.2	**73.0**±1.3	76.4±1.2	**78.2**±0.6

TABLE 4.1: Test accuracy (%) of PACS and VLCS based on training-domain validation hyper-parameter tuning. "+" mark indicates that it trained by our DOMAIN+.

4.4.2 Implementation Details

Efficient Influence Calculation. Considering the computational challenges in Eq. (4.5), we used the Second-order Stochastic estimation technique for the liner-time approximation based on implicit Hessian-vector products (HVPs) [88, 141]. In particular, we run it three times, 1,000 steps each, and average the results as the influence in Eq. (4.5).

Parameter Settings. Following the settings in DOMAINBED, we used pre-trained ResNet-50 [142] as the backbone for all methods on all datasets and optimize all models using Adam [130]. We followed the default hyper-parameter setting in DOMAINBED, where the batch size is 32 and the learning rate is 5×10^{-5}, and we tuned the threshold α based on the best performance on the validation set.

4.4.3 Quantitative and Qualitative Analysis

We show the effectiveness of our DOMAIN+ by the following Q&A. **Q1.** *How does* DOMAIN+ *improve DG methods?*

A1. Our main results are shown in Table 4.1 and Table 4.2. Compared to the original DG methods, our DOMAIN+ consistently and significantly improves most of the settings. In specific, we improve the averaged performance over the four datasets

Algorithm	OfficeHome					TerraIncognita				
	A	C	P	R	Avg	L100	L38	L43	L46	Avg
ERM [139]	59.5±1.8	52.3±1.3	73.9±1.3	75.6±1.0	65.3±0.4	49.7±2.9	43.3±1.6	56.1±2.1	35.3±2.8	46.1±1.8
DRO [74]	58.8±1.4	52.9±1.5	74.5±1.1	75.5±0.7	65.4±0.7	48.9±3.4	41.1±3.9	**57.5**±0.5	37.3±1.9	46.2±1.0
Mixup [103]	61.4±1.4	53.9±1.7	76.1±0.7	77.1±0.6	67.1±0.5	54.7±4.5	44.1±3.8	55.1±3.2	31.1±3.7	46.2±2.7
MLDG [70]	57.2±0.6	51.4±1.9	73.1±1.0	74.8±0.9	64.1±0.4	50.4±3.8	37.5±4.0	52.6±3.9	34.1±3.6	43.6±1.9
MMD [76]	58.4±1.3	53.4±0.7	73.9±0.8	75.2±0.6	65.2±0.6	48.8±3.2	40.9±3.1	54.2±1.8	36.6±4.1	45.1±2.1
ANDMask [78]	55.8±1.3	50.5±1.6	73.2±0.7	75.0±0.7	63.6±0.4	44.6±3.8	40.5±1.3	53.6±2.1	37.0±3.0	43.9±1.5
SagNet [140]	59.1±1.6	52.4±2.5	74.6±0.9	75.3±0.8	65.3±0.4	50.6±3.6	44.0±2.3	54.8±1.3	31.4±3.9	45.2±2.7
IRM [4]	58.5±1.0	52.0±1.3	73.5±1.5	74.8±0.9	64.7±0.7	53.5±4.4	41.8±3.6	55.6±1.7	37.7±4.2	47.2±1.4
IRM+	60.5±1.2	52.9±1.4	75.2±1.2	76.2±0.4	66.2±0.4	54.8±4.2	**45.4**±3.3	56.3±2.0	38.1±0.9	**48.6**±1.3
CORAL [32]	63.0±1.2	55.3±0.9	76.0±0.5	76.8±0.9	67.8±0.4	51.9±2.0	41.1±2.8	52.4±3.4	37.3±2.8	45.7±2.0
CORAL+	**63.3**±0.9	**56.4**±0.9	**76.6**±1.1	**78.4**±0.2	**68.7**±0.4	**55.5**±2.4	44.1±3.3	56.0±3.1	37.8±3.0	48.4±2.3
Fish [77]	59.0±1.4	52.4±1.8	73.7±0.7	74.5±0.6	64.9±0.8	49.9±2.0	41.7±1.7	54.5±1.6	37.9±3.5	46.0±0.9
Fish+	59.9±1.5	53.0±1.4	75.0±0.7	75.9±1.1	65.9±0.9	53.6±2.2	44.7±2.2	56.2±1.2	**38.2**±3.5	48.2±1.9

TABLE 4.2: Test accuracy (%) of OfficeHome and TerraIncognita based on training-domain validation hyper-parameter tuning. "+" mark indicates that it trained by our DOMAIN+.

of IRM, CORAL, and Fish by 1.3%, 1.6%, and 1.3%, respectively. In particular, we have achieved the largest improvement on the TerraIncognita dataset, where the baselines perform worse than other datasets. One possible reason is that this dataset may contain more spurious invariance (since its only domain variance is the latitude of the camera, there may be more shared features across domains than that in PACS), which weakens the baselines, and thus our DOMAIN+ for removing the bias caused by spurious invariance plays a more essential role in improvement.

Q2. *How does* DOMAIN+ *perform compared to SOTAs?*

A2. Compared to the original SOTAs, especially the ERM implemented by DOMAINBED [12], the DG methods equipped with our DOMAIN+ realize new SOTAs in each set of each dataset. Specifically, we improve the average SOTA performance of the four datasets by 1.1%, 0.8%, 0.9%, and 1.4%, respectively. Noteworthy, some original DG methods cannot even beat ERM. However, after applying our DOMAIN+, they outperform ERM in most cases. This demonstrates that by removing the bias induced by the spurious invariance, we effectively promote the potential invariance ability of these DG methods.

To further show the effect of DOMAIN+ in feature learning, in Figure 4.6, we visualized the extracted features of IRM (Top) and IRM+ (Bottom). At the top, we find that some samples belonging to the same class (points with the same color) are not perfectly clustered together, which leads to incorrect predictions. However,

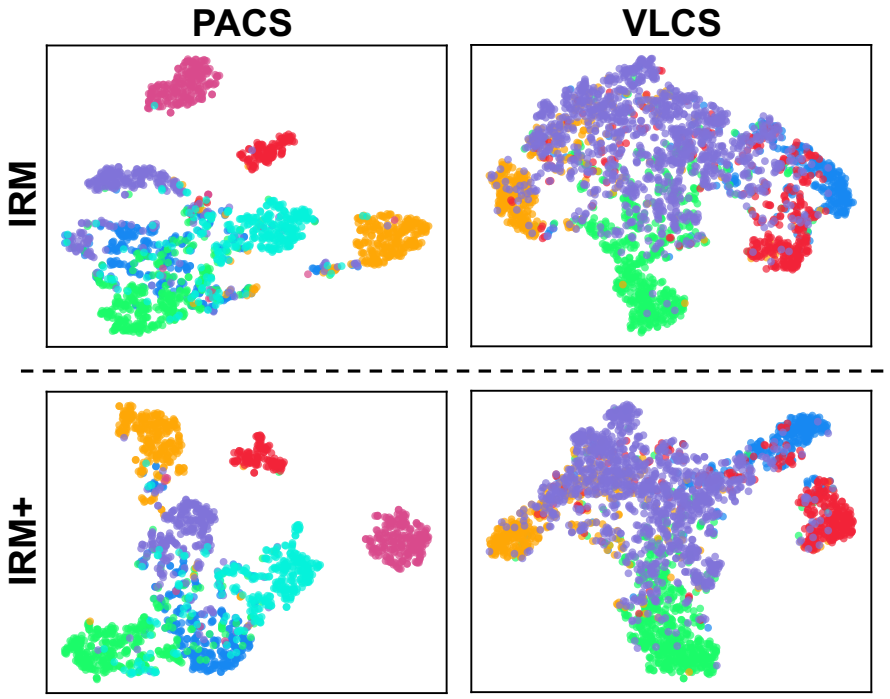


FIGURE 4.6: t-SNE [1] visualization of the features of test samples extracted by IRM and IRM+ (IRM with our DOMAIN+). We trained the model on the default three domains on each dataset. Different colors denote different classes.

Algorithm	**A**	**C**	**P**	**S**	**Avg**
IRM	85.7±2.1	79.8±1.6	95.8±0.4	78.0±1.7	84.8±0.4
IRM+ (Random)	84.3±1.4	79.8±1.5	96.2±0.9	74.4±6.0	83.6±1.8
IRM+ (Loss)	85.1±3.0	78.6±1.6	96.3±0.6	75.5±2.7	83.9±1.5
IRM+ (Ours)	85.9±2.8	81.1±0.3	96.7±0.7	78.7±1.4	85.6±0.7

TABLE 4.3: Ablations on influence. "Random", "Loss", and "Ours" denote different sample selection methods.

at the bottom, the features are clustered with much less confusion, indicating our DOMAIN+ helps IRM learn better domain-invariant features by removing the bias.

Q3. *Why do we prefer influence over training loss?*

A3. Table 4.3 shows that the new domain chosen by the influence helps DG methods achieve better performance, *i.e.*, better invariance, compared to the samples selected by the training loss. The reasons are two-fold. First, influence is a better measure of "rarity". As shown in Figure 4.4, most samples have influence values similar to 0, indicating that they are the dominant samples, while the dominant

Algorithm	A	C	P	S	Avg
IRM	85.7±2.1	79.8±1.6	95.8±0.4	78.0±1.7	84.8±0.4
IRM+	85.9±2.8	81.1±0.3	96.7±0.7	78.7±1.4	85.6±0.7
IRM w/o $\mathcal{D}^+$	79.3±3.1	70.8±2.0	92.3±1.4	74.4±1.7	79.2±0.8
IRM++	84.7±1.9	78.4±2.3	96.7±0.5	74.4±1.8	83.6±0.4
IRM-zero	84.9±2.9	80.9±1.3	95.5±0.3	78.4±2.5	85.0±1.3

TABLE 4.4: Experimental results on further exploration of DOMAIN+, where IRM++ denotes re-training IRM with DOMAIN++, IRM w/o $\mathcal{D}^+$ denotes re-training IRM without $\mathcal{D}^+$, and IRM-zero denotes no domain label is provided.

counterpart of training loss is difficult to identify because its numerical curve is not as sharp as the influence values. Second, the training loss only focuses on hard samples that do not well fit the model but do not necessarily have spurious invariance. As shown in Figure 4.5, the selected samples by training loss are far from the sample center, which means that its selection only focuses on eccentric training samples, *e.g.*, noisy samples. However, the selection of influence is more dispersed, which means that rare samples are indeed confounded by the majority distribution—spurious correlation (invariance) is identified.

Q4. *How about just training without rare samples?*

A4. Differs from the traditional influence-based methods [99, 101], which treat rare samples as "harmful" and simply discard them from training, removing rare samples will cause DG methods to learn more spurious invariance and thus reducing performance, just like the results shown in Table 4.4. We perform more experiments to demonstrate the necessity of constructing the new domain. As illustrated in Figure 4.7, we observe a decrease in invariance loss after removing rare samples, but an increase in testing loss, indicating that the model learns more spurious invariance without rare samples. In contrast, when trained on DOMAIN+, the simultaneous drop of both testing loss and invariance loss demonstrates that constructing a new domain through rare samples is necessary.

Q5. *More splits? No domain labels?*

A5. To explore the potential of DOMAIN+, we propose DOMAIN++, which applies DOMAIN+ on top of an existing split domains by DOMAIN+. Our experimental results are listed in Table 4.4. We find that compared to DOMAIN+, DOMAIN++

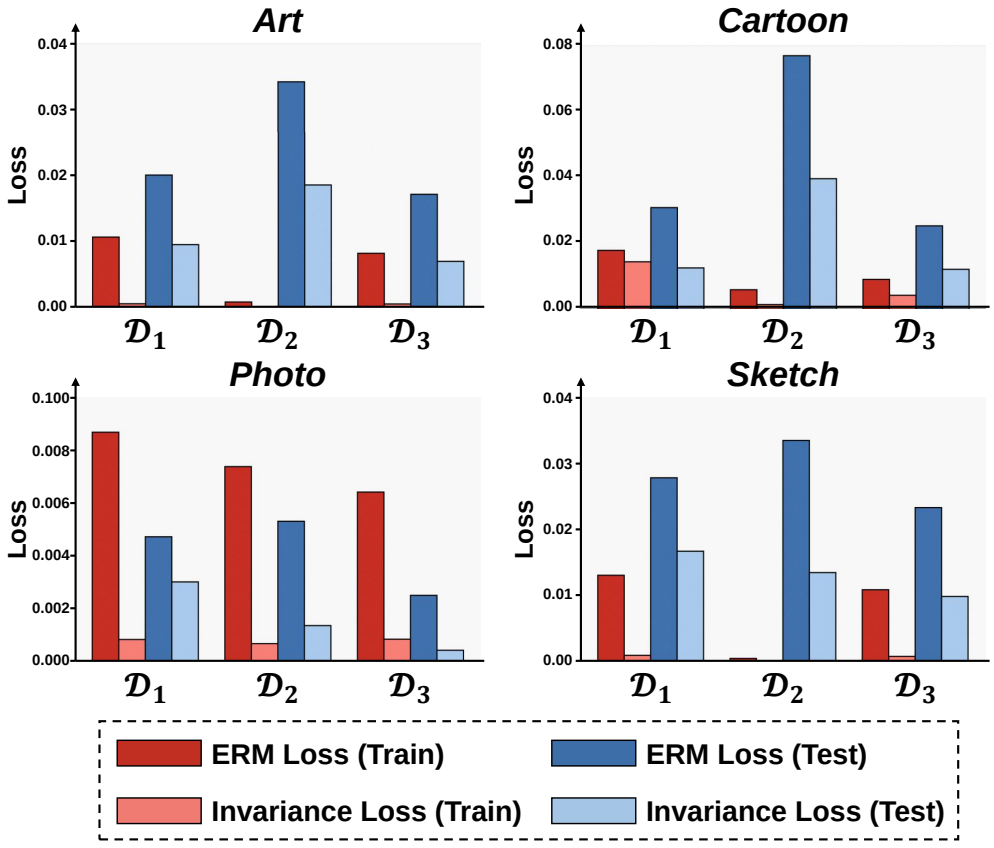


FIGURE 4.7: The training/testing ERM/Invariance loss for IRM on PACS with different setups of training domains, where $\mathcal{D}_1$ denotes the original training dataset $\mathcal{D}$, $\mathcal{D}_2$ denotes $\{\mathcal{D}_k \setminus \mathcal{D}^+\}_{k=1}^K$, and $\mathcal{D}_3$ denotes $\{\mathcal{D}_k \setminus \mathcal{D}^+\}_{k=1}^K \cup \{\mathcal{D}^+\}$, *i.e.*our DOMAIN+.

cannot further improve the performance, which is even worse than IRM. This suggests that our influence selects sufficient rare samples without spurious invariance, and further selection may introduce unexpected approximation error as the influence estimation is essentially an approximation.

We also implement DOMAIN-ZERO, when there are no domain labels. We first randomly split the training data into two domains and apply our DOMAIN+ to create a new domain. Then we can implement DG methods to learn the invariance. In Table 4.4, when we apply DOMAIN-ZERO on PACS, we still follow the conventional setting but do not use the domain labels. The improvements of IRM-zero (IRM with DOMAIN-ZERO) compared to the original IRM shows the potential future of our influenced-based domain splitting method on more tasks.

Chapter 5

OOD Generalization with no Additional Annotations[1]

5.1 Case Study: Debiasing

Different from Visual Dialog, where multiple modalities are given and the history bias and user bias can be analyzed from the underlying causal graph between them, and Domain Generalization, where multiple domains and domain labels are given and thus domain bias can be removed by the invariance methods, in this chapter, we will introduce the most challenging and common OOD Generalization camp, OOD Generalization with no Additional Annotations, and focus on the specific case, Debiasing. It is challenging because no additional annotations are given in the training to help us find the bias, and it is common because besides Debiasing [2, 82], many other OOD tasks without additional annotations also fall into this camp. In this chapter, we use the word "debiasing" to denote the traditional Implicit Debiasing setting, and we will introduce how to find the context bias in Debiasing and propose methods to remove it.

[1]The main content of the first part in this chapter is submitted to *CVPR 2023* as **Jiaxin Qi**, Qianru Sun, Xian-Sheng Hua, and Hanwang Zhang. A Two-stage Method for Training Unbiased Models. *Under Review.* and the second part is published as **Jiaxin Qi**, Kaihua Tang, Qianru Sun, Xian-Sheng Hua, and Hanwang Zhang. Class Is Invariant to Context and Vice Versa: On Learning Invariance for Out-Of-Distribution Generalization. *ECCV 2022.*

The reason for deep models to learn the context[2] bias in Debiasing settings is easy to find: they are good at data fitting by their powerful learning ability [13], and thus any data bias that is beneficial to reduce the training loss will also be encoded as class representations to help models to classify, leading to biased models that fail in testing whose distribution is different from training, *i.e.*, the Debiasing settings. For example, if most training `0` digits are in `red` and `1` digits are in `green`, then a model will recklessly capture the `color` as the discriminative representations, and thus may misclassify a `red 1` as `0` [2]. In practice, we always demand unbiased models learned from data because any real-world training data can be considered as biased compared to the unknown testing distribution—the grand challenge of Out-of-Distribution Generalization [11, 12, 14–18].

In this chapter, we propose two algorithms: TWO and IRMCon, where the first one is a simple framework to improve the current re-weighting methods in Debiasing and the second one is to disentangle context features for better implementation of the current re-weighting methods in Debiasing and Domain Generalization without domain annotations, which is also a debiasing setting and different from the settings we have discussed in Chapter 4, where the domain annotations are necessary. Now, we will delve into our algorithms in the following two sections, respectively.

5.2 Our Algorithm: TWO

5.2.1 Motivation

Under the biased training set, perhaps an effective way to learn an "always unbiased" model is to collect absolutely unbiased training data about anything with everything, *e.g.*, if we have any digit written in every possible color, font, stroke, *etc.*, we can train an unbiased digit classifier without considering color, font, and stroke. However, such data collection is prohibitively impossible, instead, most debiasing methods resort to simulating such a process by modifying the biased training distribution into an unbiased one, such as sample re-weighting/-sampling [2, 6], adversarial training [143, 144], and augmentation [6, 83]. However, as shown in Figure 5.1, when the testing distribution is the same as training, those methods are

[2]Note that the word "context" denotes any class-agnostic attribute such as color, texture, and background.

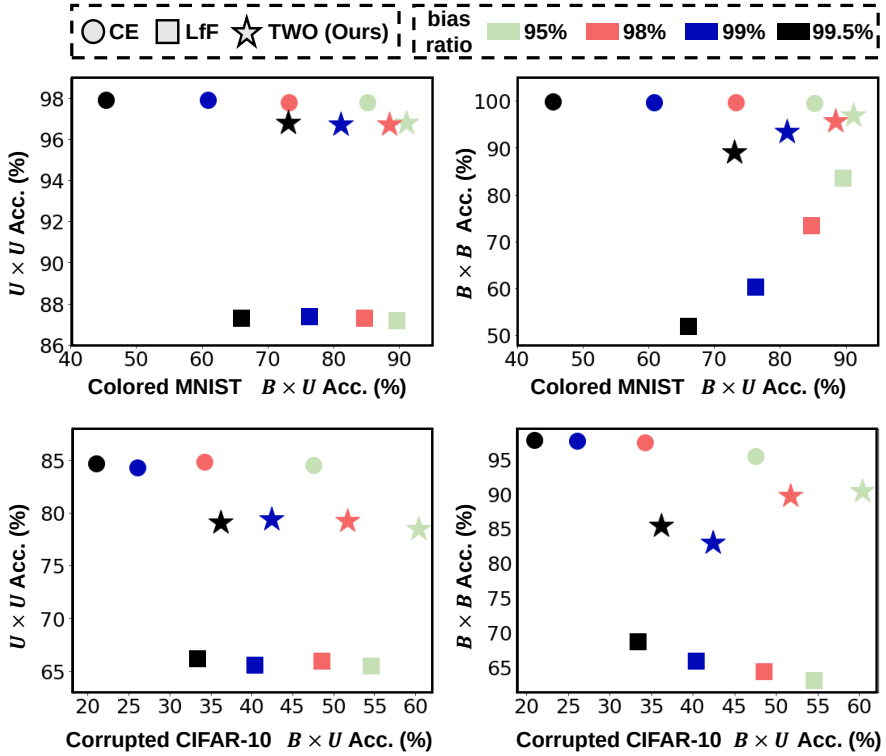


FIGURE 5.1: Comparisons on Unbiased/Biased Train $(U/B) \times U/B$ Test. The same color denotes methods tested under the same setting, and the point closer to the top right corner is better. CE achieves high accuracies on $U \times U$ and $B \times B$, but fails in $B \times U$. The one-stage re-weighting methods, such as LfF [2], are only good at $B \times U$, but fail in $U \times U$ and $B \times B$. We achieve the best performance on $B \times U$ without much sacrifice on $U \times U$ and $B \times B$. See more results in Table 5.1& 5.2.

even worse than the biased counterpart—this is clearly far from the ideal absolute unbiasedness. The reason is that such training distribution modification is essentially equivalent to duplicating the training distribution outliers (a.k.a. hard samples) to counter the biased attributes encoded in training, *e.g.*, the dominating `color` in one digit class. However, as such duplication introduces no sample diversity at all, it will also duplicate other inherently unbiased attributes, *e.g.*, digit `shape`, and thus, unfortunately, make them biased, *e.g.*, balancing the biased `color` will make the originally unbiased `shape` biased. We call such side effects as introducing the undesired "anti-training bias" bias, for which, we will provide an in-depth analysis in Section 5.2.3 and the illustrative example in Figure 5.2.

To this end, we present TWO: a simple yet effective two-stage debiasing method for training unbiased models. TWO does not only outperform existing methods on the conventional debiasing settings but also achieves a consistent performance gain

regardless of training and testing bias (Figure 5.1), showing a promising potential for training a truly unbiased model agnostic to testing distributions.

Stage-1: traditional Cross-Entropy (CE) loss training. Note that this is the key difference from existing methods, which mistakenly believe that such biased training is harmful as it only captures the bias, and hence they abandon this stage or only use it as a bias model [2]. In contrast, we show that the bias-sensitive Stage-1 indeed learns features invariant to the relatively unbiased attributes—such invariance is also invariant to the subsequent re-balancing stage, hence mitigating the "anti-training bias" bias.

Stage-2: Supervised Contrastive-regularized CE (SCCE) training only on the balanced samples selected by using a bias model [2]. Note that such balanced training is equivalent to re-weighting/sampling because the non-selected samples can be considered as down-weighted/sampled. We show that the supervised contrastive regularization [85] can learn unbiased features and why they cannot be used in Stage-1 (Section 5.2.3.4).

To evaluate the testing-agnostic unbiasedness of Two, in Section 5.2.4.1, we propose to conduct four unbiased (U) and biased (B) cross-evaluations: (U, B) training distribution $\times$ (U, B) testing distribution, and measure the performance consistency by using the harmonic mean of the four testing accuracies. The motivation is that a truly unbiased model should capture the bias-invariant (or causal) features regardless of training and testing distributions. In the debiasing experiments, we show that Two achieves the best harmonic mean on four benchmarks: Colored MNIST, Corrupted CIFAR-10, BAR, and BFFHQ.

5.2.2 Implementations

Input: Biased training data $\mathcal{D} = \{(x, y)\}$, where x is a sample and y is its associated one-hot label vector whose non-zero value index denotes the ground-truth class index. The classification model f is randomly initialized. Note that different from some debiasing methods such as Rubi [145] and EnD [146], which require the annotations of the biased attributes, we do not require such additional supervision and hence our method is more general.

Output: Unbiased classification model f.

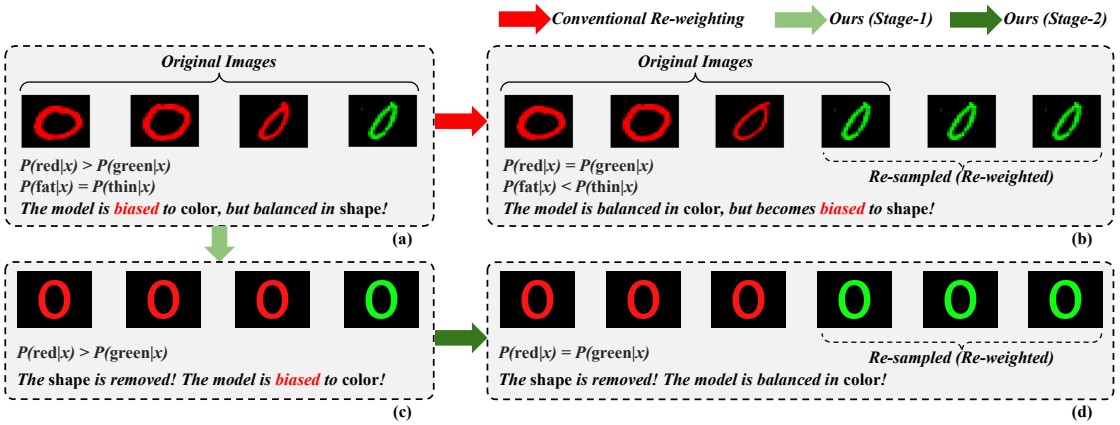


FIGURE 5.2: Illustrations for the conventional re-weighting method and our two-stage re-weighting method. (a) Original images, where `shape` is balanced but `color` is imbalanced, (b) the re-sampled (re-weighted) images by conventional re-weighting method, where `color` is balanced but `shape` becomes imbalanced, (c) images without `shape` after our Stage-1, (d) the re-sampled (re-weighted) images by our two-stage re-weighting method, where `color` is balanced.

Stage-1: Training f by CE loss on $\mathcal{D}$. We adopt the conventional Cross-Entropy (CE) loss:

$$\mathcal{L}_{\text{CE}}(f, \mathcal{D}) = \frac{1}{|\mathcal{D}|} \sum_{(x,y)\in\mathcal{D}} -y \cdot \log p(x), \tag{5.1}$$

where $p(x) = \frac{\exp(f(x))}{\sum_j \exp(f(x)_j)}$ is the softmax class confidence vector of x, $f(x)_j$ is the j-th dimension of the class logits, and "$\cdot$" is the dot product. In Section 5.2.3.1, we justify why the above CE loss is biased and why we should decouple it from the next unbiased training.

Stage-2: Training f by SCCE loss on balanced set $\mathcal{B}$. Although this stage is essentially similar to the re-weighting/re-sampling [2, 6], we take a different approach to fine-tune f on a balanced subset $\mathcal{B} \subset \mathcal{D}$, which can be considered as a 0/1 hard re-weighting (subject to a threshold) for the samples with high confidence for a wrong class:

$$\mathcal{B} = \{(x, y) | \max(p_b(x) \odot (1 - y)) > \alpha, \forall (x, y) \in \mathcal{D}\}, \tag{5.2}$$

where $p_b(x)$ is the softmax class confidence vector from a pre-trained bias model f_b, "$\odot$" is the element-wise multiplication, max selects the maximum value of the input vector, and $\alpha \in [0, 1)$ is a pre-defined threshold. Note that we call $\mathcal{B}$ " balanced"

as it consists of false-positive samples that are eccentric to the imbalanced training distribution.

We follow LfF [2] to train f_b by Generalized Cross-Entropy (GCE) loss on $\mathcal{D}$:

$$\mathcal{L}_{\text{GCE}}(f_b, \mathcal{D}) = \frac{1}{|\mathcal{D}|} \sum_{(x,y)\in\mathcal{D}} y \cdot \frac{1 - p_b(x)^q}{q}, \tag{5.3}$$

where $q \in (0, 1]$ is a temperature that controls the degree of bias amplification and $\lim_{q\to 0} \frac{1-p^q}{q} = -\log p$, *i.e.*, CE is the limit of GCE. As compared to CE, GCE amplifies the gradient of samples with high probability $p_b(x)$, encouraging f_b to quickly converge to those biased "easy-to-learn" samples.

After collecting $\mathcal{B}$, we fine-tune f on $\mathcal{B}$ by the Supervised Contrastive-regularized CE (SCCE) loss:

$$\mathcal{L}_{\text{SCCE}}(f, \mathcal{B}) = \mathcal{L}_{\text{CE}}(f, \mathcal{B}) + \beta \frac{1}{|\mathcal{B}|} \sum_{(x,y)\in\mathcal{B}} -\log\left(\frac{1}{|\mathcal{B}_i|} \frac{\sum_{x_n\in\mathcal{B}_i/x} \exp(g(x)\cdot g(x_n))}{\sum_{x_m\in\mathcal{O}/x} \exp(g(x)\cdot g(x_m))}\right), \tag{5.4}$$

where $\beta > 0$ is a trade-off hyperparameter, $\mathcal{B}_i = \{(x_i, y_i) | y_i = y, \forall (x_i, y_i) \in \mathcal{B}\}$ denotes the i-th class set, f is composed by a feature backbone g and a classification head h, *i.e.*, $f = g \circ h$, which is initialized by Stage-1, and thus inherits the learned knowledge from Stage-1. We use Supervised Contrastive loss (SC) [85] to encourage the similarity between intra-class features and penalize the similarity between inter-class features learned by g. Therefore, as compared to CE that only focuses on inter-class discrimination, SC also encourages g to map intra-class samples to a common class vector. In Section 5.2.3.4, we show why SCCE is only applicable to Stage-2.

5.2.3 Justification

Context Bias in Causal Theory: As we have discussed in Section 3.2, we take a causal view to analyze the context bias, induced by the confounder between the input image and the output class label: We denote random variables X, Y, and Z as a sample, class label, and hidden attributes, respectively. We show that the biased distribution of Z introduces bias, *e.g.*, if there are more `red 0` than `green 0` in the training set, Z = `red` can be the bias for `0`. An ideal unbiased model

should perform consistently well under any testing Z distribution, and we can say that the model captures the causality between X and Y [147–152]. In particular, we use Pearl's causality to formulate the unbiased predictor $P(Y|do(X))$ [19]:

$$P(Y|do(X)) = \sum\nolimits_z P(Y|X,z)P(z), \tag{5.5}$$

where $do(X)$ is a notation for the causal intervention of X, denoting that the prediction of Y is solely based on the intervention of input X but not other variables. Its calculation on the right-hand-side of Eq. (5.5) means that different from the conventional likelihood $P(Y|X)$ in Eq. (5.6) below, to achieve unbiased prediction, we need to put X in the context of any $Z = z$, which is independent on the observational distribution of X, *i.e.*, $P(z|X) = P(z)$. Please see Section 3.4 for a graphical explanation of $P(Y|do(X))$ and its connections to unbiased statistical inference.

5.2.3.1 CE is overall biased

Minimizing CE loss is equivalent to maximize the posterior $P(Y|X)$ [153]:

$$P(Y|X) = \sum\nolimits_z P(Y|X,z)P(z|X). \tag{5.6}$$

We use Figure 5.2(a) as an illustrative example to show why CE is biased. Y is `0` and Z is `color` that has two values {`green`, `red`}, and we assume that `color` has a uniform prior, *i.e.*, `green` and `red` have the same chance to generate samples. As $P(\texttt{red}|X) = 3/4$ and $P(\texttt{green}|X) = 1/4$, $P(\texttt{0}|X)$ will take more $P(\texttt{0}|X,\texttt{red})$ into account as this term has higher weight than $P(\texttt{0}|X,\texttt{green})$ according to Eq. (5.6). In contrast, $P(\texttt{0}|do(X))$ will equally treat $P(\texttt{0}|X,\texttt{green})$ and $P(\texttt{0}|X,\texttt{red})$ because they have the same weight, $P(\texttt{green}) = P(\texttt{red}) = 1/2$, based on Eq. (5.5). Therefore, $P(\texttt{0}|X) \neq P(\texttt{0}|do(X))$ and $P(\texttt{0}|X)$ is biased to $P(\texttt{0}|X,\texttt{red})$. Especially, when $P(\texttt{red}|X) \gg P(\texttt{green}|X)$, *i.e.*, $P(\texttt{red}|X) \approx 1$ and $P(\texttt{green}|X) \approx 0$, $P(\texttt{0}|X) \approx P(\texttt{0}|X,\texttt{red}) \approx P(\texttt{0}|\texttt{red})$, which means $P(\texttt{0}|X)$ is totally biased to `red`.

5.2.3.2 CE is partly unbiased

Without loss of generality, we assume that Z can be split into two conditionally independent attribute *sets* $Z = (Z_1, Z_2)$, $P(Z_1, Z_2|X) = P(Z_1|X)P(Z_2|X)$, where the elements in Z_1 are balanced to X, *i.e.*, $Z_1|X$ is conditional uniformly distributed, and the elements in Z_2 are imbalanced to X, *i.e.*, $Z_2|X$ is not conditional uniformly distributed. We have:

$$P(Y|X) = \sum\nolimits_{(z_1,z_2)} P(Y|X, z_1, z_2)P(z_1)P(z_2|X), \tag{5.7}$$

where the derivations are based on: Z_1 and Z_2 are conditionally independent $P(z_1, z_2|X) = P(z_1|X)P(z_2|X)$, and Z_1 is balanced to X, *i.e.*, $P(z_1|X) = P(z_1)$.

By marginalizing over Z_1, we can re-write Eq. (5.7) as:

$$P(Y|X) = \sum\nolimits_{z_2} P_{\bar{Z}_1}(Y|X, z_2)P(z_2|X), \tag{5.8}$$

where $P_{\bar{Z}_1}(Y|X, z_2) := \sum_{z_1} P(Y|X, z_1, z_2)P(z_1)$ denotes the unbiased prediction w.r.t. Z_1. Therefore, $P(Y|X)$ is unbiased to Z_1 but biased to Z_2.

As illustrated in Figure 5.2(a), Z_1 is `shape` which is balanced to X and Z_2 is `color` which is imbalanced to X. Because $P(\texttt{fat}|X) = P(\texttt{thin}|X) = 1/2$, $P(\texttt{0}|X)$ takes $P(\texttt{0}|X, \texttt{fat}, \texttt{color})$ and $P(\texttt{0}|X, \texttt{thin}, \texttt{color})$ into the same account, and hence $P(\texttt{0}|X)$ is unbiased to `shape`. To illustrate the unbiasedness, we use a standard-font `0`, *i.e.*, the `0` model is *invariant* to `shape`.

5.2.3.3 One-stage re-weighting is still biased

Re-weighting method [2, 6, 84] is to achieve $P(Y|do(X))$ by balancing the biased $P(Z|X)$ in $P(Y|X)$. Followed by Eq. (5.8), we have

$$P(Y|do(X)) = \sum\nolimits_{z_2} P_{\bar{Z}_1}(Y|X, z_2)P(z_2|X)P(X)w(z_2, X), \tag{5.9}$$

where $w(z_2, X) = \frac{P(z_2)}{P(X)P(z_2|X)} = \frac{1}{P(X|z_2)}$ is the weight (a.k.a. propensity score [19, 107, 108]). In this paper, $w(z_2, X)$ is implemented as in Eq. (5.2). Such weight essentially duplicates the samples $X \sim P(z_2|X)P(X)$ without increasing the diversity of Z_1, making $P(Z_1|X) \neq P(Z_1)$, *i.e.*, Z_1 is no longer balanced to X, and

Bias Ratio	10%	95.0%				98.0%				99.0%				99.5%			
Settings	$U\times U$	$U\times B$	$B\times U$	$B\times B$	HM	$U\times B$	$B\times U$	$B\times B$	HM	$U\times B$	$B\times U$	$B\times B$	HM	$U\times B$	$B\times U$	$B\times B$	HM
CE*	**97.8**	**97.9**	85.0	99.1	94.6	**97.9**	73.0	99.5	90.5	**97.9**	56.7	99.5	83.1	**97.9**	41.1	99.7	73.0
CE	97.6	97.8	85.2	**99.5**	94.6	97.8	73.3	**99.7**	90.6	**97.9**	60.9	99.7	85.2	**97.9**	45.4	**99.8**	76.1
Rebias [82]	97.4	97.6	85.0	99.4	94.5	97.6	73.5	**99.7**	90.6	97.6	61.4	**99.8**	85.4	97.6	45.7	**99.8**	76.3
LfF [2]	85.2	87.2	89.6	83.7	86.4	87.3	84.7	73.5	82.3	87.4	76.3	60.3	75.7	87.3	66.0	52.0	69.5
JTT [81]	95.5	95.6	87.1	98.1	93.9	95.7	79.0	98.2	91.4	95.6	65.7	98.5	86.4	95.7	55.5	99.2	81.6
Feat-Aug [6]	90.1	89.7	89.0	96.5	91.2	89.7	85.2	78.5	85.6	89.7	77.3	64.3	78.9	89.7	70.2	54.8	73.1
SoftCon [84]	95.9	96.6	85.3	99.3	93.9	96.6	74.9	99.5	90.5	96.6	64.6	99.5	86.4	96.6	48.7	99.7	77.9
Two w/o SC	96.1	96.7	90.0	96.9	94.8	96.6	86.7	95.6	93.6	96.7	78.7	93.4	90.5	96.7	71.7	89.0	87.1
Two w/ SC	96.3	96.8	**91.1**	96.9	**95.2**	96.7	**88.5**	95.6	**94.1**	96.7	**81.1**	93.4	**91.4**	96.8	**73.1**	89.0	**87.7**

TABLE 5.1: Comparisons on Colored MNIST. As $U \times U$ is unchanged across different settings, we only show it in the first column. The columns with gray background are the conventional debiasing evaluation $B \times U$. CE* denotes no augmentation. We reproduce the compared methods using their officially released codes under the same experimental settings.

we consider it as the "anti-training bias" bias. Therefore, the pitfall of one-stage re-weighting methods lies in the fact that we haven't obtained $P_{\bar{Z}_1}(Y|X, z_2)$ yet in one-stage methods and thus re-weighting may introduce the "anti-training bias" bias:

$$P(Y|do(X)) \neq \sum_{(z_1,z_2)} P(Y|X, z_1, z_2)P(z_1|X)P(z_2|X)P(X)w(z_2, X), \tag{5.10}$$

where the right-hand side denotes the one-stage training. Interestingly, the above analysis shows a potential direction for one-stage debiasing: generating samples that do not break the balance of Z_1. However, it may require the challenging disentanglement of Z_1 and Z_2 to generate counterfactual samples [154]. As shown in Figure 5.2(b), Z_1 is `shape` and Z_2 is `color`. We find that `color` is imbalanced to X, and `shape` is balanced to X. To balance `color`, the weight estimator will assign weight 3 for (`green`, X) and weight 1 for (`red`, X), illustrated as copying `green 0` for three times. After re-weighting, $P(\texttt{green}|X) = P(\texttt{red}|X)$, *i.e.*, `color` is balanced. But, $P(\texttt{thin}|X) > P(\texttt{fat}|X)$, *i.e.*, `shape` becomes imbalanced to X. This imbalance will make the model biased to `shape`.

5.2.3.4 Our two-stage re-weighting (Two) is unbiased

Thanks to the above justification, if we adopt the two-stage method, we can ensure $P_{\bar{Z}_1}(Y|X, z_2)$ before re-weighting according to z_2, and hence Eq. (5.9) is valid to achieve the unbiased $P(Y|do(X))$. As shown in Figure 5.2(c), after Stage-1, the

model is unbiased to `shape` (Z_1), *i.e.*, it treats the hand-written `0` with different `shape` as the standard-font `0` invariant to `shape`. Then, we can safely re-weight the imbalanced `color` (Z_2) by duplicating the standard-font `green 0` for three times to achieve $P(z_2 = \texttt{green}|X) = P(z_2 = \texttt{red}|X)$. After re-weighting, we can achieve the unbiasedness for all Z, *i.e.*, `shape` (Z_1) is removed after Stage-1 and `color` (Z_2) is removed after Stage-2.

Moreover, **why do we only apply Supervised Contrastive (SC) loss in Stage-2 but not Stage-1**? Recall that SC loss encourages the similarity of sample pairs in the same class. So, when X is biased to Z, the prevailing Z is more likely to be misrecognized as the unbiased class feature. For example, in Figure 5.2(a), when `red 0` is dominant, SC loss encourages the model to learn the feature of `red` into the class feature to increase the similarities between sample pairs, *i.e.*, SC loss makes model biased to `red`. In contrast, as shown in Figure 5.2(d), when `color` is balanced, SC loss will not encourage the model to learn `red` or `green` into the class feature. Therefore, we only apply SC loss in Stage-2, when Z is balanced.

5.2.4 Experiments

We first introduce our experimental setups, including the details of the benchmarks, evaluation metrics, comparing methods, and our implementations in Section 5.2.4.1. Then, we demonstrate the effectiveness of our TWO with quantitative and qualitative analysis, and we discuss the failure cases and the future improvements in Section 5.2.4.2.

5.2.4.1 Experimental Setups

Datasets. We followed LfF [2] and Feat-Aug [6] to use two synthetic datasets, Colored MNIST [2] and Corrupted CIFAR-10 [2], and two real-world datasets, Biased Action Recognition (BAR) [2] and Biased Flickr-Faces-HQ (BFFHQ) [6]. By associating a specific bias to a class, we can construct an unbiased/biased training set and unbiased/biased testing set with different bias ratios.

For the synthetic ones, we controlled the bias in the generation process. For Colored MNIST, we selected 10 different colors and assign each color randomly to a unique class with a specific ratio to generate the color-biased datasets. For

Bias Ratio	10%	95.0%				98.0%				99.0%				99.5%			
Setting	$U\times U$	$U\times B$	$B\times U$	$B\times B$	HM	$U\times B$	$B\times U$	$B\times B$	HM	$U\times B$	$B\times U$	$B\times B$	HM	$U\times B$	$B\times U$	$B\times B$	HM
CE*	73.2	74.0	32.2	92.3	57.9	74.3	23.7	93.7	50.1	73.9	19.4	95.2	44.8	74.2	17.5	92.7	42.1
CE	**82.9**	**84.5**	47.6	**95.4**	72.2	**84.8**	34.3	**97.4**	63.2	**84.3**	26.1	97.6	55.2	**84.7**	21.0	**97.7**	48.9
Rebias [82]	82.7	83.8	47.3	94.9	71.8	83.7	33.1	97.1	62.0	83.5	25.0	**97.9**	53.8	83.5	20.9	81.8	47.5
LfF [2]	64.6	65.5	54.6	63.1	61.6	66.0	48.5	64.4	59.9	65.6	40.4	65.9	56.6	66.2	33.4	68.7	53.2
JTT [81]	79.9	81.0	50.6	94.1	72.4	81.5	36.6	96.3	64.0	81.1	28.2	96.6	56.6	81.2	24.0	97.6	52.1
Feat-Aug [6]	73.0	73.7	54.1	68.4	66.3	73.8	45.2	62.9	61.3	73.5	36.9	62.9	56.9	74.0	34.2	69.9	56.6
SoftCon [84]	79.9	81.0	51.1	94.7	72.7	81.3	37.4	96.3	64.6	81.3	27.9	96.2	56.3	81.1	23.3	97.1	51.2
Two w/o SC	78.3	78.5	56.8	90.4	73.8	79.2	47.2	89.7	69.3	79.4	41.9	82.9	65.2	79.1	33.8	85.4	59.9
Two w/ SC	78.3	78.5	**60.4**	90.4	**75.3**	79.2	**51.7**	89.7	**71.6**	79.4	**42.4**	82.9	**65.6**	79.1	**36.2**	85.4	**61.8**

TABLE 5.2: Comparisons on Colored CIFAR-10. As $U \times U$ is unchanged across different settings, we only show it in the first column. The columns with gray background are the conventional debiasing evaluation $B \times U$. CE* denotes no augmentation. We reproduce the compared methods using their officially released codes under the same experimental settings

	Bias Ratio	CE*	CE	Rebias	LfF	JTT	Feat-Aug	SoftCon	Two w/o SC	Two w/ SC
BAR	95.0%	51.1±1.0	53.6±0.8	55.1±0.5	52.7±1.8	52.6±2.1	53.5±1.1	54.9±0.8	56.5±3.0	**60.2±0.6**
	99.0%	37.9±2.1	41.6±3.4	43.7±0.5	43.0±3.6	42.2±2.0	42.4±2.9	45.1±2.2	44.1±1.2	**48.1±2.6**
BFFHQ	99.5%	70.6±0.2	70.8±1.8	70.0±3.8	71.5±0.2	73.0±1.7	71.4±0.5	71.6±0.6	72.9±1.8	**73.8±0.3**

TABLE 5.3: Comparisons on BAR and BFFHQ. Methods are evaluated by the traditional debiasing evaluation (*i.e.*, $B \times U$) as other settings are not applied.

example, in the 95.0% biased training set, 95.0% of images with class label `0` are `green`. The same protocol is applied to generate the Corrupted CIFAR-10 dataset. We selected 10 corruptions, {`Gaussian`, `Gaussian Blur`, `Defocus Blur`, `Frost`, `Saturate`, `Pixelate`, `Elastic`, `Impulse`, `Brightness`, `Contrast`}, and used each corruption to corrupt a certain class in CIFAR-10 by some ratios to generate corruption-biased datasets. We selected the ratio in {99.5%, 99.0%, 98.0%, 95.0%, 10.0%} to create unbiased/biased training sets and unbiased/biased testing sets. Note that, as there are 10 classes in the two datasets, the 10.0% bias ratio denotes 10 colors (corruptions) are uniformly distributed in each class, *i.e.*, unbiased setting. For the real-world datasets, we instead select images with specific bias. BAR [2] contains six kinds of action-place bias, *e.g.*, the action `fishing` always happens on the `water surface`. We chose a bias ratio in {99.0%, 95.0%} to construct biased training sets. For BFFHQ face datasets, we set the attribute `age` as the bias for class `gender` (*i.e.*, most of the `females` are `young` and `males` are `old`) for the ratio of 99.5% to construct its biased training set.

Evaluation Metrics. We used the unbiased/biased training sets and unbiased/biased testing sets to design 4 cross-bias evaluation settings. 1) $U \times U$: Unbiased training and Unbiased testing, 2) $U \times B$: Unbiased training and Biased testing, 3) $B \times U$: Biased training and Unbiased testing, 4) $B \times B$: Biased training and Biased testing. As we believe that the true debiasing method should be robust under any combination of training and testing distributions, we used the harmonic mean (denoted as HM) of the accuracies on the four settings as the overall metric: $HM = 4/(1/Acc_{U\times U}+1/Acc_{U\times B}+1/Acc_{B\times U}+1/Acc_{B\times B})$. Note that, limited by the data generation process, we couldn't apply the four settings on BAR and BFFHQ. Therefore, we used the traditional evaluation, *i.e.*, $B \times U$. As in [2], we took three independent runs and used the mean accuracy to calculate HM.

Comparing Methods. We compared our Two with CE baseline, CE with augmentations, and 5 SOTA methods which also do not need bias annotations: Rebias [82], LfF [2], JTT [81], Feat-Aug [6] and SoftCon [84]. Note that we carefully selected the augmentations to ignore the leakage of bias, *i.e.*, we didn't use color-related augmentations in Colored MNIST and didn't use corruption-related augmentations in Corrupted CIFAR-10.

Implementation Details. For a fair comparison, we reproduced all of the methods by using their officially released codes under the same settings: We used MLP with three hidden layers for Colored MNIST, and ResNet-18 for Corrupted CIFAR-10, BAR, and BFFHQ. We used the cosine scheduled SGD optimizer with the initial learning rate as 0.1 for Colored MNIST and 0.05 for Corrupted CIFAR-10, Adam [130] optimizer for two real-world datasets with a learning rate of 0.001. We set the batch size as 256 and training 100 epochs for Colored MNIST, Corrupted CIFAR-10, batch size 256 and training 50 epochs for BFFHQ, batch size as 64, and training epochs as 150 for BAR. We conduct all experiments using Tesla V100 GPUs.

5.2.4.2 Quantitative and Qualitative Analysis

Overall Results. We show the results on the synthetic datasets in Table 5.1 and Table 5.2, and real-world datasets in Table 5.3. Overall, for both harmonic mean and traditional evaluation, our Two achieves the best performance on all of the benchmarks and leads a clear margin to the compared methods.

	Colored MNIST				Corrupted CIFAR-10			
	95.0%	98.0%	99.0%	99.5%	95.0%	98.0%	99.0%	99.5%
w/o Stage-1	92.8	90.1	85.7	80.0	26.5	25.2	24.5	24.2
LfF as Stage-1	92.4	90.3	85.3	82.0	52.3	51.7	48.7	46.7
Two	95.2	94.1	91.4	87.7	75.3	71.6	65.6	61.8

Table 5.4: Ablations for Stage-1. "w/o Stage-1" denotes that we train the Stage-2 from scratch, "LfF as Stage-1" denotes that we replace CE with LfF [2] in Stage-1. The values are the harmonic mean.

Specifically, 1) on the synthetic datasets, in Table 5.1 and Table 5.2, following the order from low bias ratio (95.0%) to high bias ratio (99.5%), when we evaluate the methods by harmonic mean, our Two improves the SOTA by 0.6%, 2.7%, 5.0%, 6.1% on Colored MNIST, and 2.6%, 7.0%, 8.7%, 5.2% on Corrupted CIFAR-10. As for the traditional evaluation, $B \times U$, our Two improves the SOTA by 1.5%, 3.3%, 3.8%, 2.9% on Colored MNIST, and 5.8%, 3.2%, 2.0%, 2.0% on Corrupted CIFAR-10. 2) on the real-world datasets in Table 5.3, evaluated by $B \times U$, we can improve the SOTA by 5.1% and 3.0% on BAR {95.0%, 99.0%}, respectively, and 2.2% on BFFHQ. These improvements demonstrate that our two-stage re-weighting is indeed more unbiased than the one-stage ones.

Verification for Our Justifications. From Table 5.1 and Table 5.2 we find some experimental verifications for our justifications in Section 5.2.3.

1) CE (or its augmentation variant) dramatically loses its performance under the $B \times U$ setting. The reason is CE is overall biased, as we discussed in Section 5.2.3.1, under the higher bias ratio, CE will learn more bias.

2) The conventional one-stage re-weighting methods, such as LfF [2] and Feat-aug[6], fail in the cross-bias settings except for $B \times U$, leading to an unsatisfied harmonic mean. The reason is that the one-stage re-weighting introduces the "anti-training bias" bias, as we discussed in Section 5.2.3.3, which makes the model still biased.

3) Our two-stage re-weighting achieves the best harmonic mean, and we also achieve the best accuracy on the traditional evaluation $B \times U$, compared to the one-stage re-weighting methods. The reason is that our two-stage re-weighting is theoretically superior to the conventional one-stage re-weighting, as we discussed in Section 5.2.3.4, and we can achieve unbiasedness.

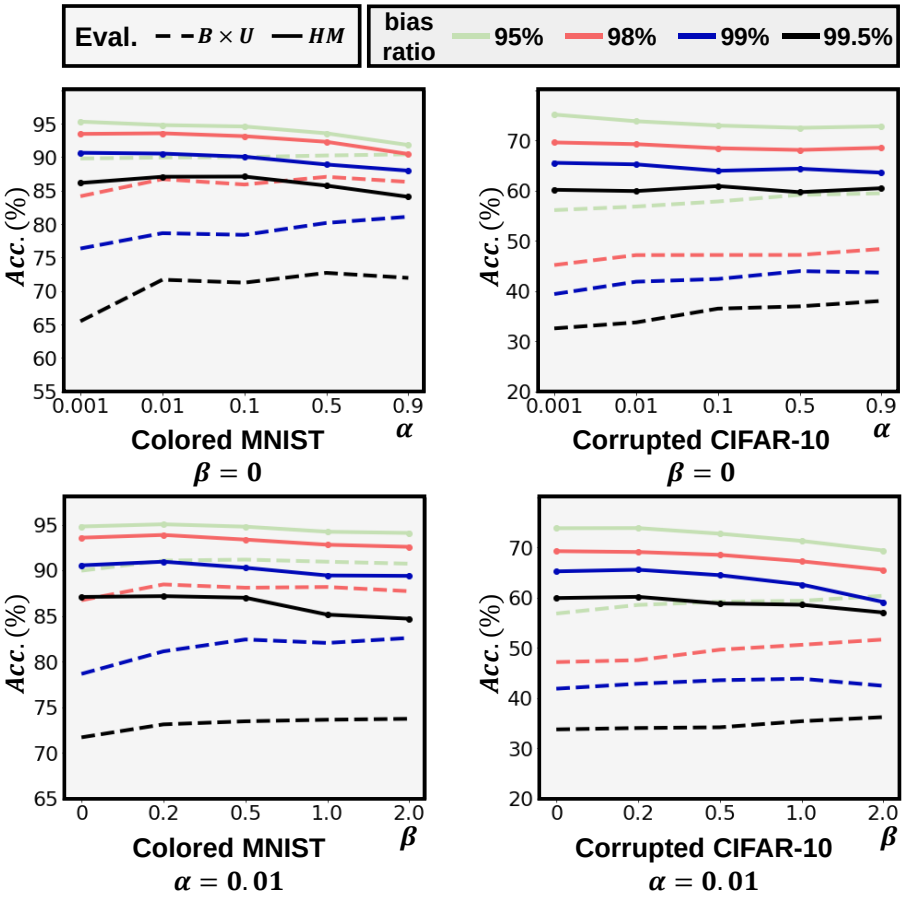


Figure 5.3: Ablations for α in Eq. (5.2) and β in Eq. (5.4) on Colored MNIST and Corrupted CIFAR-10. Green, red, blue, and black lines denote the four bias ratios 95.0%, 98.0%, 99.0%, and 99.5%, respectively. We illustrate the accuracies in $B \times U$ (dashed lines) and HM (solid lines).

Ablation Studies. 1) Ablations for α and β. As illustrated in Figure 5.3, the left two ablations justify that our method is not sensitive to α. The right two ablations for β justify that SC helps Stage-2. Specifically, we find that the accuracy improved with the increase of α, the reason is that we select samples based on the bias model—the higher confidence predicted by the bias model, the more accurate for the bias estimation. Therefore, increasing the threshold can filter out the inaccurate bias predictions and select a more balanced set for Stage-2.

2) Ablations for our Stage-1. In Table 5.4, to justify the necessity of our Stage-1 we conducted two ablations, eliminating Stage-1 and replacing CE with LfF unbiased model in Stage-1. The result shows that both will downplay Two in each setting. Therefore, the unbiasedness learned by Stage-1 is necessary for our two-stage re-weighting. It is worth noting that the ablation "w/o Stage-1" is the

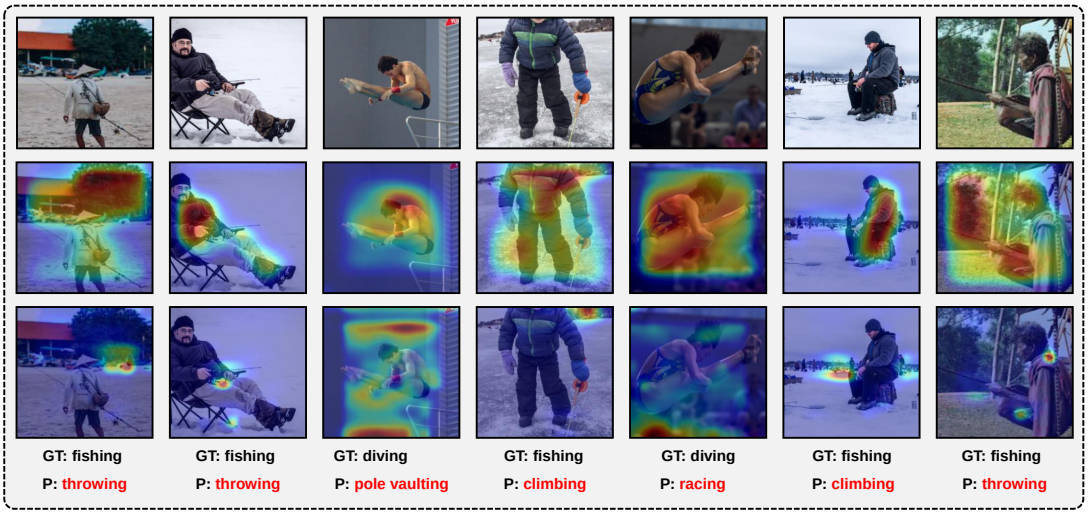


FIGURE 5.4: GradCAM [3] visualization of our failure cases due to the inaccurate bias model. Top: input test images; Middle: visualization of captured bias by the bias model; Bottom: visualization of our unbiased model. The model is trained on the 95.0% biased BAR training set. GT: ground-truth label; P: predicted label by our TWO.

vanilla implementation of one-stage re-weighting. But, most of the conventional one-stage re-weighting methods, such as LfF [2], perform much better than "w/o Stage-1". The reason is that they coincidentally get the partial help of our TWO: In the implementation of LfF [2], they train the bias model and unbiased model from scratch at the same time, and in the early training stage, when the bias model is immature and cannot capture any bias, it will assign 1 to each sample, and the unbiased model conducts CE training, like our Stage-1. However, as their "Stage-1" is insufficient and cannot be fully invariant to all the balanced attributes Z, they are still inferior to our full implementation of the two-stage re-weighting.

Failures Cases. As there is no bias annotation, existing methods learn the bias feature by the proxy classification task. However, such a task inherently entangles the "good" but undesired class feature and the "bad" but desired bias (or context) feature. To avoid learning the class feature, they elaborately design the feature backbone or training process. For example, Rebias [82] uses the 1×1 convolution backbone to learn bias features, as they believe that the smaller convolution kernel captures more context bias than class. LfF [2] and we apply GCE loss (Eq. (5.3)) to learn the "easier" bias feature, as the bias feature is "easier" to learn than the class feature. But, these technical designs cannot essentially remove the class feature, which can always help the loss minimization. As illustrated in Figure 5.4, our debiased model may fail, *i.e.*, attending to the background to predict human actions.

Accordingly, we can see that the bias model captures the class feature, as shown in the middle row, which attends to foreground human actions. This inaccurate bias estimation will cause the unsatisfied selection of $\mathcal{B}$, and eventually makes our Stage-2 biased. To this end, a possible improvement is to replace the proxy task with a class-agnostic objective to fundamentally disentangle class features and bias features, which is similar to our algorithm that we will introduce in Section 5.3.

5.3 Our Algorithm: IRMCon

5.3.1 Motivation

In this section, different from proposing a better training framework, *i.e.*, our TWO, we focus on the weight estimation problem in traditional re-weighting-based methods to propose a better weight estimation algorithm. As we have discussed in Section 5.2, to collect a good training dataset, a criterion is to ensure that the samples for each class are as diverse and plentiful as possible, and meanwhile, the diversity between classes is as evenly distributed as possible [128, 155]. For example, the "dog" class should contain images of dogs with different colors, types, and backgrounds. As shown in Figure 5.5 (a), under such dataset, an Empirical Risk Minimization (ERM) objective [20], *e.g.*, usually the softmax cross-entropy loss [120], can keep the class feature by ignoring the inter-class features due to they cannot help the model to decrease the training loss, *e.g.*, the background cannot help the model to identify the action. And this can be summarized into the common principle for classification:

Principle 1. Class is invariant to context.

For example, a "dog" image is always a dog regardless of its type, color, and background, which is the practical meaning of "is invariant to".

When we get an unbiased model trained by a balanced training set, even given testing samples whose contexts are out-of-distribution compared to the training, it can still classify correctly because it already removes the context bias and only learns the context-invariant, *i.e.*, class/causal feature [9, 16, 156], and thus the model has generalization ability [157–159]. However, due to limitations in data collection, in practice, real-world datasets are far from balanced and learning class

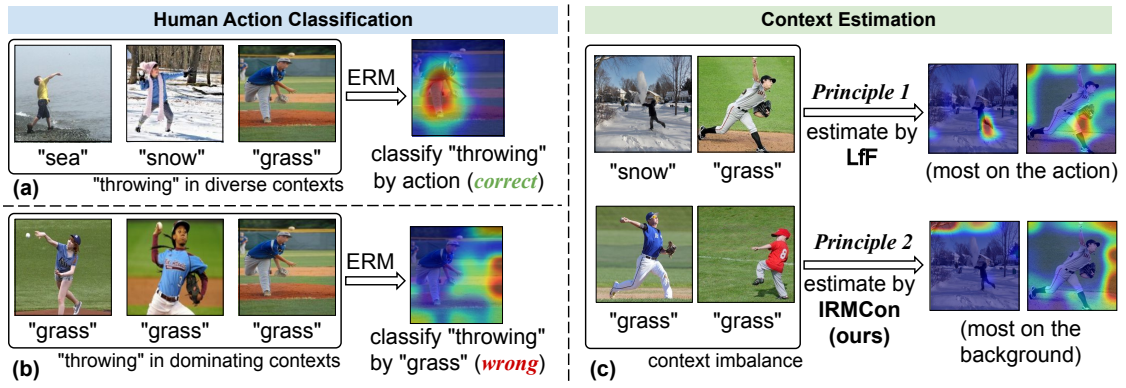


FIGURE 5.5: GradCAM [3] visualizations of the learning of ERM. (a): By using ERM, if the context is diverse and balanced in the training images of a class, the model trained by such dataset will focus on the human's action to predict the class. (b): If one context dominates one class image in the training, the model will learn context into the class feature, which is used for classification, *e.g.*, the background "grass" is for classification. (c): The traditional context estimation [2] based on Principle 1, which is biased to class,*e.g.*, the context estimation model focuses on human action "throwing" to estimate the context, while our IRMCon based on Principle 2 estimates better context, *e.g.*, focusing on the background to estimate context.

invariance on unbalanced datasets is challenging. As shown in Figure 5.5 (b), in the training images of class "throwing", if the context "grass" dominates, the spurious correlation between "grass" and "throwing" will be learned by the model, inducing context bias which hurt the model performance, because the model will use the "grass" to predict "throwing". As a result, the problem for debiasing is the imbalanced context distribution in the training.

To realize a balanced context training set, we need to estimate the context first for the training set. The current methods for context estimation or context bias estimation fall into two categories. First, they annotate the context directly [4, 7], as shown in Figure 5.6 (c). The case is discussed in Chapter 4 but the annotations are not given in most of the cases. Besides, it is not possible to fully annotate the complex context. For example, it is easy to label crude scenes as "water" and "grass", but it is difficult to distinguish their nuances further. Consequently, context monitoring is usually incomplete. Second, all of them estimate the context by the biased class prediction [2, 6, 82], as illustrated in Figure 5.6 (d). Such estimation relies on the contra-position of Principle 1, which is fundamentally an indirect context estimation.

Principle 1. (Complement) If a feature is not invariant to context, it is not class but context.

Here, the judgment of *"not invariant to context"* is achieved by using the biased prediction of the classifier, *i.e.*, if the classifier predicts incorrectly, it is due to that the class invariance has yet been achieved in the classifier. Unfortunately, as classification inference is a combined effect of class and context, it is difficult to distinguish whether the bias comes from a biased context or from unsophisticated class modeling. The reflection in the result is the incorrect context estimation mixed with class effect (see the top part of Figure 5.5 (c) and the qualitative results in the experiments). In fact, coinciding with recent findings [12, 16], we show in Section 5.3.4 that the current methods with inaccurate context estimation may even underperform the ERM baseline. In particular, if the data is less biased, such methods may catastrophically mistake context for class, this limits their applicability only in severely biased training data.

In this section, we propose a more direct and accurate context estimation method without using any additional context labels. Our inspiration comes from the other side of Principle 1:

Principle 2. Context is also invariant to class.

For example, the background "grass" is always grassy regardless of its foreground object class, such as the class "dog" or "cat".

Principle 1 implies that the success of learning class invariance is due to a changing context. Similarly, Principle 2 tells us that we can learn context invariance by changing classes, which is easier for us to achieve because the classes (taken as varying environments [4]) are already labeled and balanced, a common practice with any supervised training data, with an equal sample size for each class. In Section 5.3.3, as illustrated in Figure 5.6 (e), we propose a context estimator trained by minimizing the contrastive loss of intra-class sample similarity which is invariant to classes (based on Principle 2). In particular, the invariance is achieved by Invariant Risk Minimization (IRM) [4] with our new loss term. We call our method **IRMCon** where **Con** stands for context. Figure 5.5 (c) illustrates that our IRMCon can capture better context features. Based on IRMCon, we can simply deploy a re-weighting method, *e.g.*, [160], to generate the balancing weights for different contexts, eventually removing the context bias and achieving class invariance.

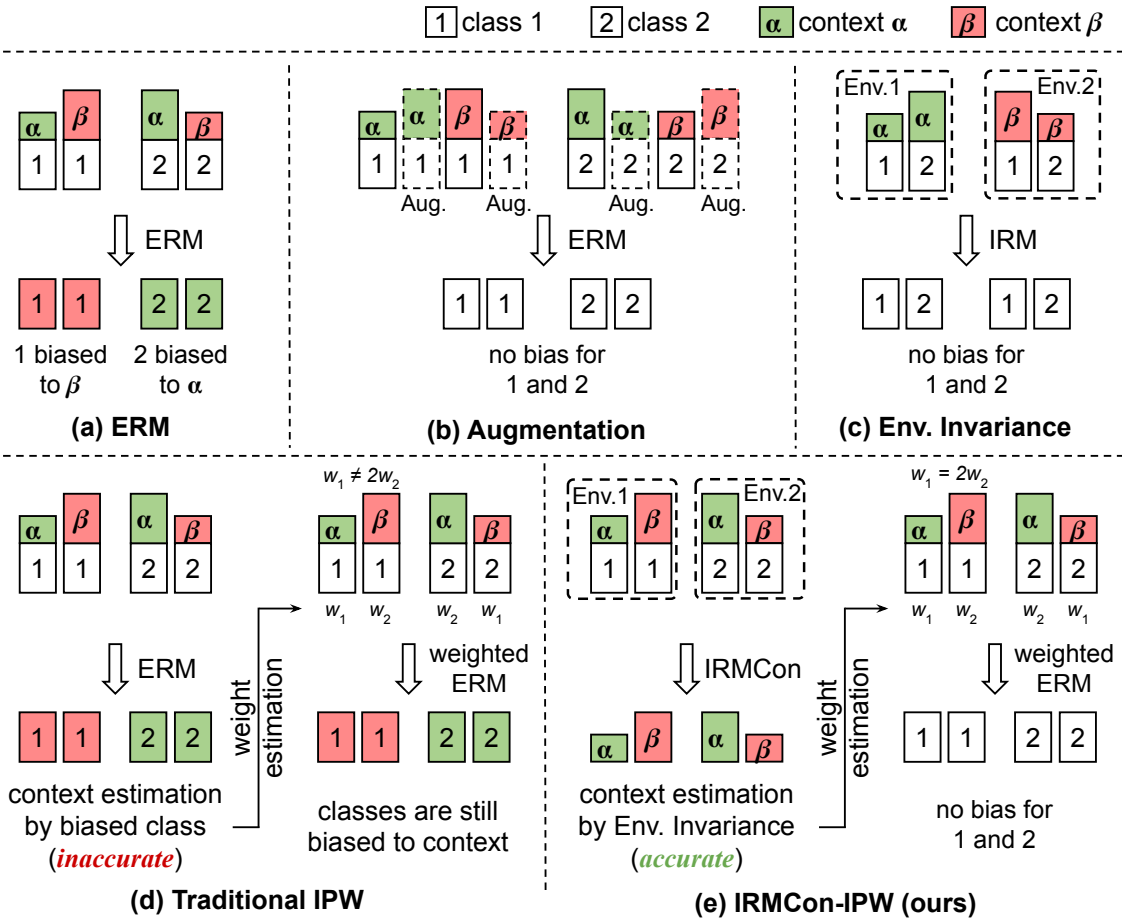


FIGURE 5.6: Illustrations of the related methods [2, 4–9]. ERM denotes the baseline methods. Others and ours are aiming for removing context bias, where the details of other methods including augmentation, invariance, and re-weighting are given in Section 2.1. We explain the components in the following: 1) The length of a context (colored) bar denotes the number of samples of that context, such as "grass", where the longer bar denotes that the context is more dominant, *e.g.*, longer "grass" bar than "sand" bar denotes more images are in "grass" background compared to "sand" background in one class. 2) A single colored bar with a class number denotes the learned class feature is biased to the prevailing context. Our algorithm IRMCon-IPW is based on the Invariance methods (IRM) and the re-weighting methods (IPW), and our main contribution, compared to the traditional methods, is trying to disentangle context features not by using class objectives but by eliminating class features. We provide the theoretical justifications in Section 5.3.3 and we evaluate our algorithm in Section 5.3.4.2.

We follow DOMAINBED [12], which we have discussed in Domain Generalization Experiment 4.4, for rigorous and reproducible evaluations, including 1) a strong Empirical Risk Minimization (ERM) baseline that is used to be mistakenly poor in OOD, and 2) a fair hyper-parameter tuning validation set. Experimental results in

Section 5.3.4 demonstrate that our IRMCon can effectively learn context variance and eventually improve the context bias estimation, leading to a state-of-the-art debiasing performance.

5.3.2 Preliminaries: Invariance as Class

Model generalization in supervised learning is based on the fundamental assumption [17, 161]: any sample x is generated from the two disentangled features (or independent causal mechanisms [149]), $x = g(\mathbf{x}_c, \mathbf{x}_t)$, where $\mathbf{x}_c$ is the class feature, $\mathbf{x}_t$ is the context feature, $g(\cdot)$ is a generative function that transforms the two features in vector space to sample space (*e.g.*, pixels). In particular, the disentanglement naturally encodes the two principles. To see this for Principle 1, if we only change the context of x and obtain a new image x', we have $\mathbf{x}_c = \mathbf{x}'_c$ but $\mathbf{x}_t \neq \mathbf{x}'_t$—class is invariant to context; Principle 2 can be interpreted in a similar way. Therefore, we'd like to learn a feature extractor $\phi_c(x) = \mathbf{x}_c$ that helps the subsequent classifier to predict robustly across varying contexts.

Note that, the following two sections, *i.e.*, Section 5.3.2.1 and Section 5.3.2.2 have been discussed in Section 4.2 and we re-write them here for convenience.

5.3.2.1 Empirical Risk Minimization (ERM)

If the training data per class is balanced and diverse, *i.e.*, containing sufficient samples in different contexts, ERM has been theoretically justified that it can learn the class feature extractor $\phi_c(x)$ by minimizing a contrastive based loss such as softmax cross-entropy (CE) loss [17]:

$$\mathcal{L}_{\text{ERM}}(\phi_c, f) = \frac{1}{N}\sum_{i=1}^{N} \text{CE}(y_i, \hat{y}_i = f(\phi_c(x_i))), \tag{5.11}$$

where y_i is the ground-truth label of x_i and $\hat{y}_i$ is the predicted label by the softmax classifier $f(\cdot)$.

However, when the data is imbalanced and less diverse, ERM cannot learn $\phi_c(x) = \mathbf{x}_c$. We illustrate this in Figure 5.6 (a): if more class `1` samples contain context β than α, the resultant $\phi_c(x)$ will be biased to the prevailing context, *e.g.*, features

for classifying class `1` will be entangled with context β. To this end, augmentation-based methods [5, 162] aim to compensate for the imbalance (Figure 5.6 (b)). However, as contexts are complex, augmentation will be far from enough to compensate for all of them.

5.3.2.2 Invariant Risk Minimization (IRM)

If context annotation is available, we can use IRM [4] to learn ϕ_c by applying Principle 1 that ϕ_c should be invariant to different contexts. Compared to ERM on balanced data that achieves invariance in a passive way via random trials [163], IRM on imbalanced data adopts the active intervention, taking contexts as the environments:

$$\mathcal{L}_{\text{IRM}}(\phi_c, \theta) = \sum_{e} \frac{1}{|e|} \sum_{(x_i, y_i) \in e} \left[\text{CE}(y_i, \hat{y}_i) + \lambda \|\nabla_\theta \, \text{CE}(y_i, \hat{y}_i^\theta)\|^2 \right], \tag{5.12}$$

where $\hat{y}_i^\theta = f(\phi_c(x_i) \cdot \theta)$, e is one of the environments of the training data according to context labels, and $\lambda > 0$ is a trade-off hyper-parameter for the invariance regularization term. θ is a dummy classifier, whose gradient is not applied to update itself but to calculate the regularization term in Eq. (5.12). The minimization of both terms encourages ϕ_c in different environments close to the same baseline, *i.e.*, invariance to the environment (context) is achieved. We follow IRM [4] to set θ as 1.

As illustrated in Figure 5.6 (c), if we want to learn a common classifier that discriminates `1` and `2` in both environments, the only way is to remove the context α and β. However, it has been demonstrated by [16, 164] that the context annotation is usually incomplete and using it may even under-perform ERM.

5.3.2.3 Inverse Probability Weighting (IPW)

When context annotation is unavailable, we can estimate the context and then re-balance data according to context. We begin with the following ERM-IPW loss [117, 118]:

$$\mathcal{L}_{\text{ERM-IPW}}(\phi_c, \phi_t, f) = \frac{1}{N} \sum_{i=1}^{N} \text{CE}(y_i, \hat{y}_i = f(\phi_c(x_i))) \cdot \frac{1}{P(x_i \mid \phi_t(x_i))}. \tag{5.13}$$

Note that popular re-weighting methods are based on IPW, where the context bias is caused by a confounder, just like the bias we mentioned in causal theory 3.2, although most of them do not delve into the underlying theory. Since we have already discussed confounding effects and de-confounding methods, in this section we will not explore it redundantly and simply use the concept to denote re-weighting. We can see that the key difference between ERM-IPW and ERM is the sample-level IPW term $1/P(x_i|\phi_t(x_i))$, where $\phi_t(x) = \mathbf{x}_t$ is the context feature extractor. This IPW implies that if x is more likely associated with its context $\mathbf{x}_t$, *i.e.*, the class feature counterpart $\mathbf{x}_c$ is also more likely associated with $\mathbf{x}_t$, we should under-weight the loss because we need to discourage such a context bias.

However, the context estimation of ϕ_t is almost challenging as learning ϕ_c. Instead, a prevailing strategy is to estimate it by a biased classifier [2, 6], *e.g.*,

$$P(x|\phi_t(x)) \propto \frac{\mathrm{CE}(y, \hat{y} = f(\phi_c(x))) + \mathrm{CE}(y, \hat{y} = f_b(\phi_b(x)))}{\mathrm{CE}(y, \hat{y} = f_b(\phi_b(x)))}, \tag{5.14}$$

where ϕ_b is the bias feature extractor and f_b is the bias classifier. ϕ_b and f_b are minimized by ERM equipped with generalized cross entropy (GCE) loss [109]:

$$\mathcal{L}_{\mathrm{ERM}}(\phi_b, f_b) = \frac{1}{N}\sum_{i=1}^{N} \mathrm{GCE}(y_i, \hat{y}_i = f_b(\phi_b(x_i))), \tag{5.15}$$

where the detail form of GCE has been discussed in Eq. (5.3) and we re-write it here for convenience, $\mathrm{GCE}(y, \hat{y}) = \sum_{k=1}^{n} y_k \cdot \frac{1-\hat{y}_k^q}{q}$ is used to amplify the bias, where q is a constant, k is the index of class and n is the class number. However, the loss in Eq. (5.15) inevitably includes the effect from the class feature $\mathbf{x}_c$, due to the aforementioned assumption $x = g(\mathbf{x}_c, \mathbf{x}_t)$. In other words, such a combined effect cannot distinguish whether the bias is from class or context, resulting in inaccurate context estimation. We show the illustration in Figure 5.6 (d). Specifically, the weights are estimated from class and context, and thus still fail to obtain unbiased class features. In addition, the experimental results in Figure 5.10 (Bottom) testify that: inaccurate context estimation will severely hurt the performance, *i.e.*, fail to derive unbiased classifiers.

5.3.3 Implementations: Invariance as Context

To tackle the inaccurate context estimation of $\phi_t(x)$, we propose to apply Principle 2 as a way out. As illustrated in Figure 5.6 (e), if we consider each class as the environment, we can clearly see that the *unique* environmental change is the class that has been already labeled. This motivates us to apply IRM to learn invariance as context by removing the environment-equivariant class. The crux is how to design the contrastive-based loss—more specifically, how to modify θ and $\text{CE}(\cdot)$ in Eq. (5.12). The following is our novel solution.

We design a new contrastive loss based on the intra-class (environment) sample similarity, as follows,

$$\mathcal{L}_{ct}(\phi_t, e, \theta) = \sum_{x_i \in e} -log \frac{exp(\phi_t(x_i)^T \phi_t(\texttt{Aug}(x_i)) \cdot \theta)}{\sum_{x_i' \in e} exp(\phi_t(x_i)^T \phi_t(x_i') \cdot \theta)}, \tag{5.16}$$

where $\texttt{Aug}(\cdot)$ is the common augmentations, such as flip and Gaussian noise (used in standard contrastive losses [165–167]), e is the environment split by class, *e.g.*, under the environment e_1, any $x_i \in e_1$ has the class label $\texttt{1}$, θ is the dummy classifier, we add θ here for the convenience to introduce Eq. (5.17). The reason for using contrastive loss is that it preserves all the intrinsic features of each sample [17, 168]. Yet, without the invariance to class, $\phi_t(x) \neq \mathbf{x}_t$. Then, based on Eq. (5.12), our proposed IRMCon for learning "invariance as context" is:

$$\mathcal{L}_{\text{IRMCon}}(\phi_t, \theta) = \sum_e \frac{1}{|e|} [\mathcal{L}_{ct}(\phi_t, e, \theta) + \lambda |\nabla_\theta \mathcal{L}_{ct}(\phi_t, e, \theta)|], \tag{5.17}$$

where θ plays the same role in Eq. (5.12), to regularize ϕ_t be invariant to environments (classes). We can prove that solving Eq. (5.17) achieves $\phi_t(x) = \mathbf{x}_t$, *i.e.*, the context feature is disentangled. As demonstrated in Figure 5.8, ϕ_t can extract accurate context features. Thanks to ϕ_t, we can further improve IPW:

$$P(x|\phi_t(x)) \propto \frac{\text{CE}(y, \hat{y} = f(\phi_c(x))) + \text{CE}(y, \hat{y} = f_b(\mathbf{x}_t))}{\text{CE}(y, \hat{y} = f_b(\mathbf{x}_t))}, \tag{5.18}$$

where $\mathbf{x}_t = \phi_t(x)$. We train f_b by using GCE loss, just replacing $\phi_b(x)$ with $\mathbf{x}_t$ in Eq. (5.15). ϕ_t is trained by IRMCon and then fixed when estimating the context.

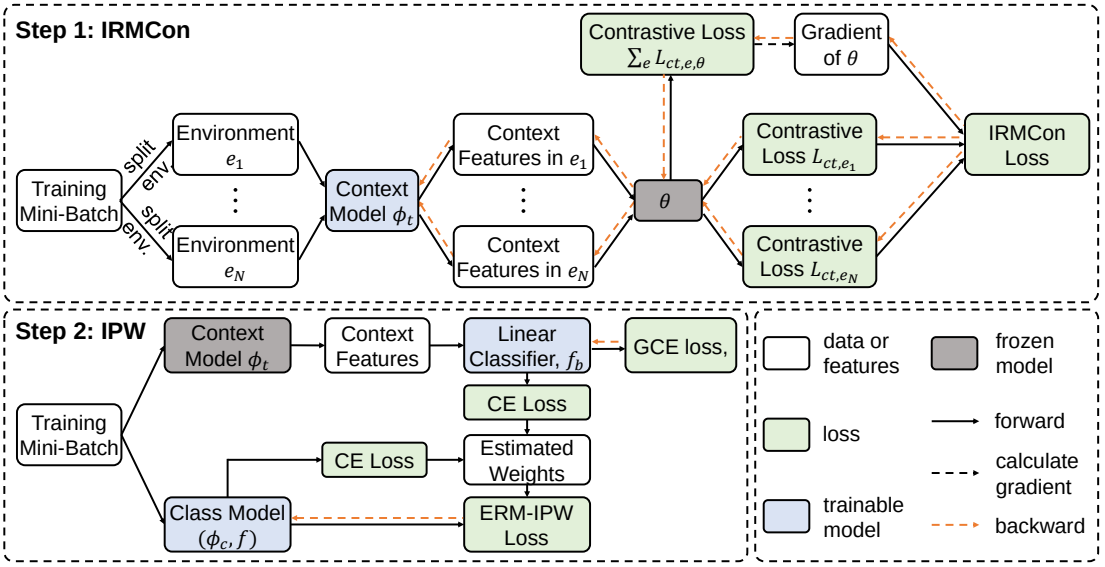


FIGURE 5.7: The training pipeline of our IRMCon-IPW. 1) "split env." denotes we split the training samples in mini-batch into subsets based on class labels, *i.e.*, samples in the class constructing one subset, forming N environments $\{e_i\}_1^N$; 2) θ is a dummy classifier, whose gradient is not for updating itself but for regularizing ϕ_t for becoming invariant to classes. 3) The black solid arrows indicate the forward calculation process and the orange dashed arrows indicate the backward propagation of the gradient.

As shown in Figure 5.9, our biased classifier can estimate more accurate weights to perform better reweighting than the traditional one We streamline the proposed IRMCon-IPW in Figure 5.7 and summarize our algorithm in Algorithm 2.

5.3.4 Experiments

We introduce the benchmarks of two debiasing tasks for removing context bias, traditional debiasing (which is similar to the experiments in Section 5.2.4) and domain generalization without domain labels (which is similar to the experiments in Section 4.4 without needing of domain labels), and our implementation details are in Section 5.3.4.1. Then, we evaluate the effectiveness of our approach based on the experimental results in Section 5.3.4.2.

5.3.4.1 Experimental Setups

Traditonal Debiasing Datasets. We follow LfF [2] to use two synthetic datasets, *Colored MNIST* and *Corrupted CIFAR-10*, and one real-world dataset, *Biased*

Algorithm 2: IRMCon-IPW

Step 1. IRMCon

Input: Training set $\{(x_i, y_i)\}_{i=1}^n$

Output: Context feature extractor ϕ_t

Randomly initialize ϕ_t;

while not converged **do**

  Sample a mini-batch from the training set;

  Split it into environments by class label;

  Update ϕ_t by IRMCon loss in Eq. (5.17).;

Step 2. IPW

Input: Training set $\{(x_i, y_i)\}_{i=1}^n$, context feature extractor ϕ_t

Output: Context invariance classifier f

Randomly initialize f_b, f, ϕ_c;

while not converged **do**

  Sample a mini-batch from the training set;

  Use frozen ϕ_t to extract context features $\mathbf{x}_t$;

  Estimate $P(x|\phi_t(x))$ in Eq. (5.18).

  Update f_b by GCE loss in Eq. (5.13). with $\mathbf{x}_t$ as input;

  Update f, ϕ_c by ERM-IPW loss in Eq. (5.13).;

Action Recognition (*BAR*) [2] for evaluation. On each dataset, we manually control the context bias ratio by generating (in synthetic datasets) or sampling (in the real-world dataset) training images.

In specific, on *Colored MNIST*, we follow LfF to generate 10 colors as 10 contexts. We connect each digit (class) with a specific color and dye them with the ratio from {99.9%, 99.8%, 99.5%, 99.0%, 98.0%, 95.0%} to construct each biased training set. In the test set, 10 colors are uniformly distributed on the samples of each class. For *Corrupted CIFAR-10*, we follow LfF to use {Saturate, Elastic, Impulse, Brightness, Contrast, Gaussian, Defocus Blur, Pixelate, Gaussian Blur, Frost} as 10 contexts. Similar to *Colored MNIST*, we generate a debiasing training set by pairing a context and a class with a ratio chosen from {99.5%, 99.0%, 98.0%, 95.0%}. In the test set, 10 contexts are uniform for each class.

The real-world dataset *BAR* contains six kinds of action-place bias, and each one is between human action and background, e.g., "throwing" always happens with the "grass" background; We choose a bias ratio in {99.0%, 95.0%}.

Domain Generalization Dataset without Domain Labels. We use *PACS* [11] to testify our method. It consists of seven object categories spanning four image

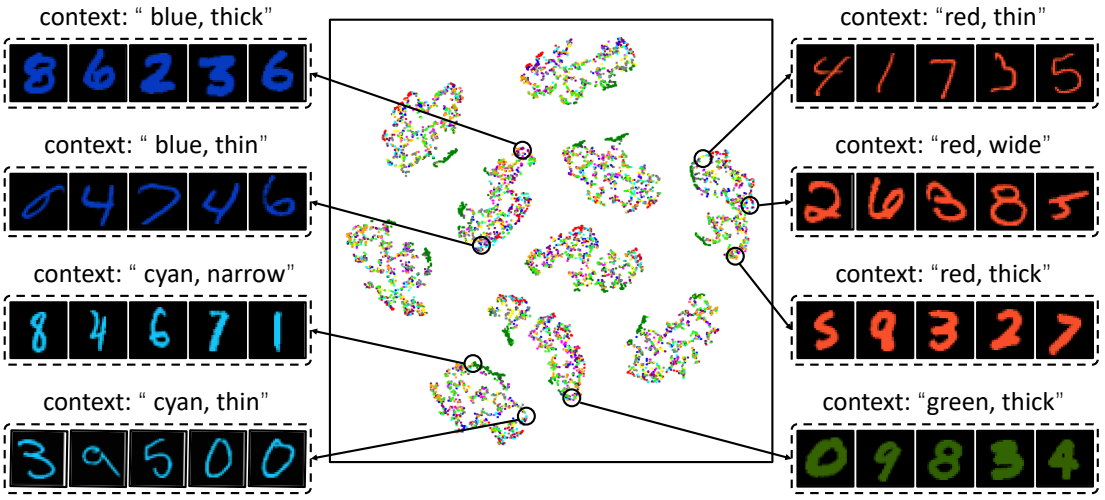


FIGURE 5.8: t-SNE [1] visualizations of our context features extracted from the *Colored MNIST* test samples, by our IRMCon model trained on the 99% biased training set. The color of the points represents their class labels and features are naturally clustered by context. As there are no context ground truths, the context labels, such as “blue, thin”, are interpreted by us.

domains: *Photo*, *Art-painting*, *Cartoon*, and *Sketch*. We follow DOMAINBED [12] to each time select three domains for training and the left one for testing. Different from the datasets setups in Chapter 4, domain labels are not given in the training, which makes Domain Generalization indeed become a debiasing problem.

Comparing Methods. As the two types of datasets have their own state-of-the-art (SOTA) methods, we compare ours with different SOTA methods in debiasing benchmark and domain generalization benchmark, respectively.

For debiasing datasets, we compare with Rebias [82], End [146], LfF [2], and Feat-Aug [6]. For the domain generalization dataset, we compare with domain-label based methods, such as DANN [80], fish [77], and TRM [169], as well as domain-label free methods, such as RSC [69] and StableNet [9]. As we claimed at the end of Section 5.3.2.1, we train all models from scratch. This makes some DG methods (*e.g.*, MMD [76] and CDANN [31]) hard to converge.

Implementation Details. We first introduce two implementation details to deal with the implementation issues we met and then provide training details.

1) *Weighted sample strategy*. This strategy is for the biased dataset. For example, with a 99.9% biased training set, all images in a mini-batch may have the same context in a class unless we can sample more than 1000 images in each class to

Dataset	Bias Ratio(%)	Methods					
		ERM	Rebias [82]	EnD* [146]	LfF [2]	Feat-Aug* [6]	IRMCon-IPW (Ours)
Colored MNIST	99.9	20.4±1.1	20.8±0.6	-	56.8±1.6	-	**66.7**±2.3
	99.8	26.4±0.4	28.3±0.9	-	68.3±1.5	-	**75.5**±1.5
	99.5	42.9±1.1	44.4±0.5	34.3±1.2	77.0±1.5	65.2±4.4	**81.0**±0.9
	99.0	59.2±0.5	58.6±0.4	49.5±2.5	82.5±1.7	81.7±2.3	**85.3**±0.3
	98.0	72.5±0.2	73.5±1.0	68.5±2.2	84.1±1.5	84.8±1.0	**88.3**±0.2
	95.0	85.7±0.5	85.5±0.5	81.2±1.4	86.8±0.5	89.7±1.1	**92.2**±0.5
Corrupted Cifar-10	99.5	22.7±0.5	22.7±0.7	22.9±0.3	26.1±0.7	30.0±0.7	**31.0**±0.6
	99.0	25.8±0.6	24.9±0.7	25.5±0.4	31.8±0.7	36.5±1.8	**37.1**±0.4
	98.0	28.7±0.1	29.1±0.7	31.3±0.4	38.9±1.0	41.8±2.3	**42.5**±1.0
	95.0	39.9±1.6	38.9±1.7	40.3±0.9	51.3±0.9	51.1±1.3	**53.8**±1.3
BAR	99.0	52.9±0.7	52.1±0.5	-	48.1±2.7	52.3±1.0	**55.3**±0.6
	95.0	65.2±1.9	65.0±1.8	-	60.6±2.6	63.5±1.5	**67.9**±0.8

TABLE 5.5: Accuracy (%) on debiasing datasets compared with SOTA methods. We reproduced the methods and averaged the results over three independent trials (mean±std). "*": For reproducing mismatch issues, performance is quoted from the original paper. "-": no report in that setting.

obtain a sample of 1 with unbiased context. To address this problem, we use the bias model from LfF [2] to learn an inaccurate context estimator and sample a relatively context-balanced mini-batch based on its inverse probability. This strategy frees us from sampling a very large batch to learn Eq. (5.16).

2) *Strategy for learning augmentation-related context.* When using contrast loss, it is difficult to learn the context associated with augmentation. To minimize contrastive loss, the model needs to learn invariance to augmentations, *i.e.*, augmentation-related features will be removed. On *Corrupted Cifar-10*, we add the classification objective in Eq. (5.15) to our IRMCon loss to train the context extractor. Please note that we use this strategy only for *Corrupted Cifar-10* because context on this dataset is dominated by augmentation-related contexts, such as 95% "car" has augmentation-related context 'Gaussian noise". For other cases, we do not need this strategy.

3) *Training details.* On the *Colored MNIST*, we use 3-layers MLPs to model ϕ_c, ϕ_b and ϕ_t. On the *Corrupted Cifar-10*, we use ResNet-18 for ϕ_c and 3-layers CNNs for ϕ_b and ϕ_t. On the *BAR* and *PACS*, we use ResNet-18 for ϕ_c, ϕ_b and ϕ_t. For optimization in debiasing datasets, we follow LfF [2] to use Adam [130] optimizer

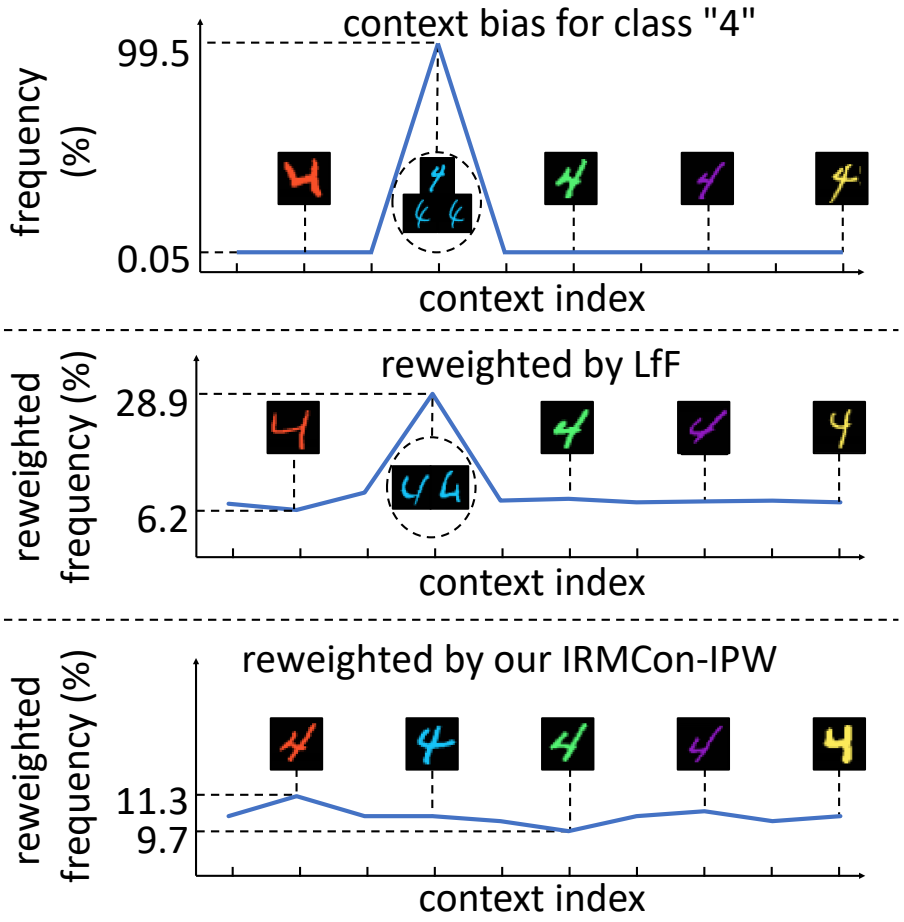


FIGURE 5.9: Illustrations of the re-weighted sample frequencies for 10 color contexts. All models are trained on the 99.5% biased *Colored MNIST* training set. The re-weighted frequency of a context indicates the normalized sum over the inverse probabilities of the samples in that context. **Top**: The distribution of biased context in the training set. **Middle**: The distribution of biased context derived by using LfF [2]. **Bottom**: Relatively balanced distribution of context obtained by using our IRMCon.

with the learning rate as 0.001. We set batch size as 256, 256 and 64 for *Colored MNIST*, *Corrupted Cifar-10* and *BAR*, respectively. We totally train 50, 50 and 250 epochs for *Colored MNIST*, *Corrupted Cifar-10* and *BAR*, respectively. For *PACS*, we apply Adam optimizer with 0.001 learning rate for 100 epochs training from scratch for all the methods.

On all datasets, we randomly split the original unbiased test set into 20% and 80% as the validation set and test set, respectively, according to DOMAINBED [12]. We selected the best model based on validation results. We averaged the results of three independent runs and reported them in the format "mean accuracy ± standard deviation".

	Methods	PACS				
		Art.	*Cartoon*	*Photo*	*Sketch*	Avg.
w/ domain supervision	IRM[4]	31.1±1.4	38.7±2.5	-	44.4±2.2	-
	DRO [74]	39.0±1.9	53.8±1.2	63.6±2.9	**62.4**±0.6	54.7
	InterMix [103]	42.2±0.5	52.8±1.9	61.0±2.4	58.4±1.0	53.6
	MLDG [70]	38.8±0.7	53.5±0.7	63.3±0.1	60.2±1.2	54.0
	DANN [80]	31.5±1.1	48.2±1.6	58.1±1.5	44.9±0.7	45.7
	V-REx [73]	33.9±1.2	40.9±1.2	-	55.1±2.9	-
	Fish [77]	**43.1**±2.1	**57.4**±0.4	**64.8**±2.7	61.1±0.8	**56.6**
	TRM [169]	41.8±1.8	54.9±0.8	-	61.3±2.3	-
w/o domain supervision	ERM	40.4±0.7	54.3±0.3	63.7±0.4	58.9±2.6	54.3
	SD [71]	39.1±0.8	54.4±1.4	61.7±3.8	51.3±3.2	51.6
	RSC [69]	41.2±2.8	49.8±6.0	58.0±1.9	53.3±4.3	50.6
	StableNet [9]	36.8±0.8	48.9±0.6	58.0±1.0	-	-
	LfF [2]	38.2±1.4	50.4±0.9	58.0±0.6	60.4±1.2	51.8
	IRMCon-IPW	**41.8**±0.5	**58.1**±1.4	**64.9**±0.7	**61.1**±2.5	**56.5**

TABLE 5.6: Accuracy (%) on the domain generalization dataset *PACS* [11]. We reproduced all the methods by the DOMAINBED [12] code base without pretraining. Results are averaged over 3 independent trials (mean±std). "-" denotes that methods fail to converge when training from scratch.

5.3.4.2 Quantitative and Qualitative Analysis

IRMCon-IPW achieves SOTA. We show our results of debiasing datasets in Table 5.5 and domain generalization dataset in Table 5.6.

1) Table 5.5 presents that our IRMCon-IPW achieves very clear margins over the related methods.

In particular, the improvements are more pronounced in the settings of higher bias rates. The possible reason for this is that when the bias ratio is higher, the samples with a "rare" context become fewer. And then re-weighting methods are more sensitive to the accuracy of context weight estimation. Therefore, accurate context estimation plays a more important role. Compared to the related methods, our IRMCon can estimate more accurate context, *i.e.*, extract high-quality context features as illustrated in Figure 5.8, and the gains are more pronounced than other methods when the context bias ratio is increased.

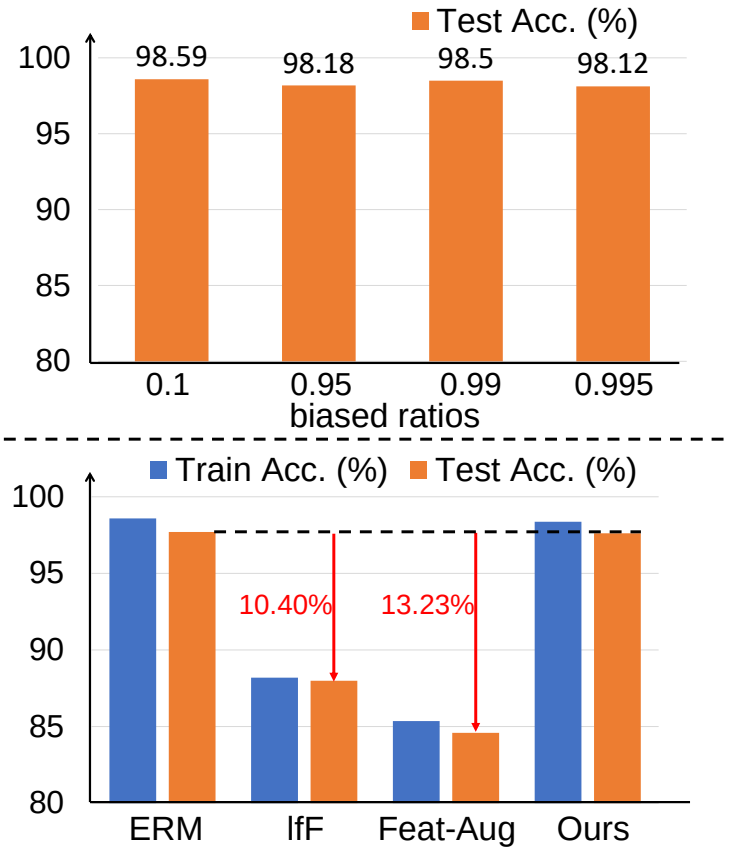


FIGURE 5.10: Accuracy (%) of models when trained on *Colored MNIST* context-balance set. **Top**: ERM is stable in test sets with varying context biases; **Bottom**: the traditional re-weighting method degrades significantly compared to ERM when trained on the contextually balanced set due to incorrect contextual estimation. Due to the correct context estimation, our IRMCon-IPW achieves comparable performance to ERM.

2) Table 5.6 shows that our method outperforms ERM on the domain generalization dataset and also achieves the best average performance among all methods without domain labels. In addition, it achieves comparable results to other domain generalization methods that require domain labels (in the upper block).

Why does ERM perform so well in most cases? On *PACS*, we follow the DOMAINBED [12] to implement a strong ERM baseline. On *BAR*, we use the strong augmentation strategy, Random Augmentation [170], which can be considered as a debiasing method as shown in Figure 5.6 (b). If we do not apply such strong augmentations, the performance of the ERM will be significantly reduced.

Why do we train models from scratch? We challenge the traditional pre-training setups in some debiasing tasks, such as Domain Generalization without

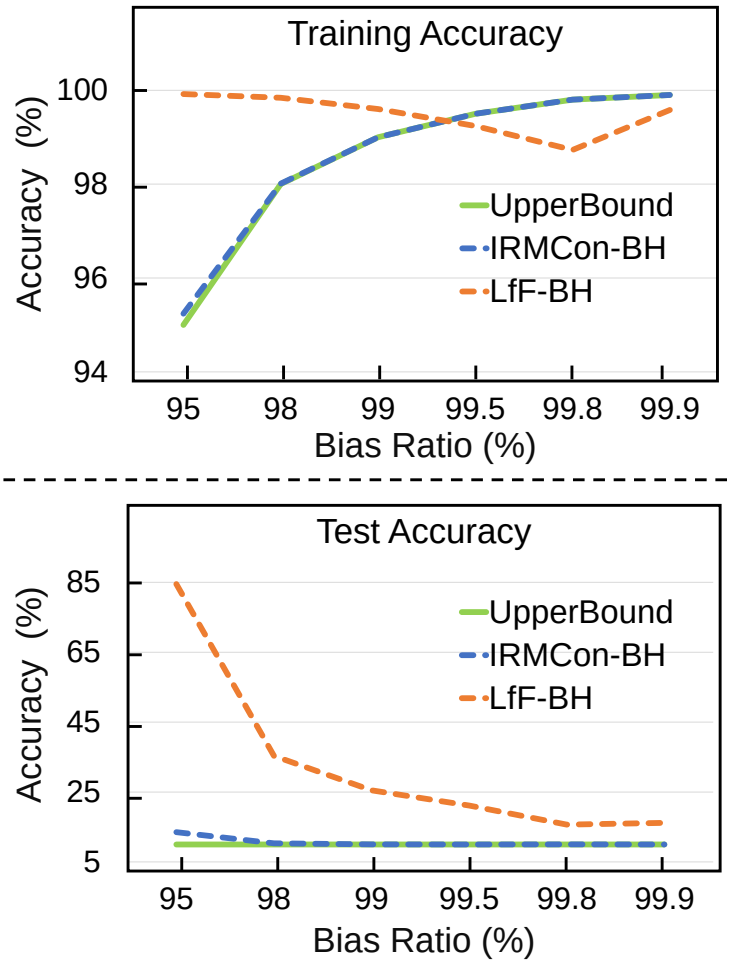


FIGURE 5.11: Comparing the bias classification head of the learned biased model of LfF [2] (LfF-BH) and of ours (IRMCon-BH) trained on *Colored MNIST* training sets with different bias ratios, where the bias classification heads (BH) intentionally use context to predict class. The figure shows that our biased learning head is almost identical to the upper bound case in the test set, random class prediction (10%).

domain labels, because we are concerned that the data or knowledge from the test set has been leaked to the model when pre-trained on large-scale image datasets. Data leakage is a common problem in pre-training settings, such as ImageNet [155] leaks to CUB [171]. Such a problem will severely destroy the validity of these task [172]. Empirically, we provide an observation in Domain Generalization to justify our challenge. In pre-training settings, ERM achieves the "impressive" 98% test accuracy [12] when *Photo* is the testing domain. This performance is significantly higher (around 20% higher) than using *Cartoon* and *Sketch* as testing. However, this is not the case if the model is not pre-trained on ImageNet, see Table 5.6, first row at the bottom, ERM method. The reason for this is that

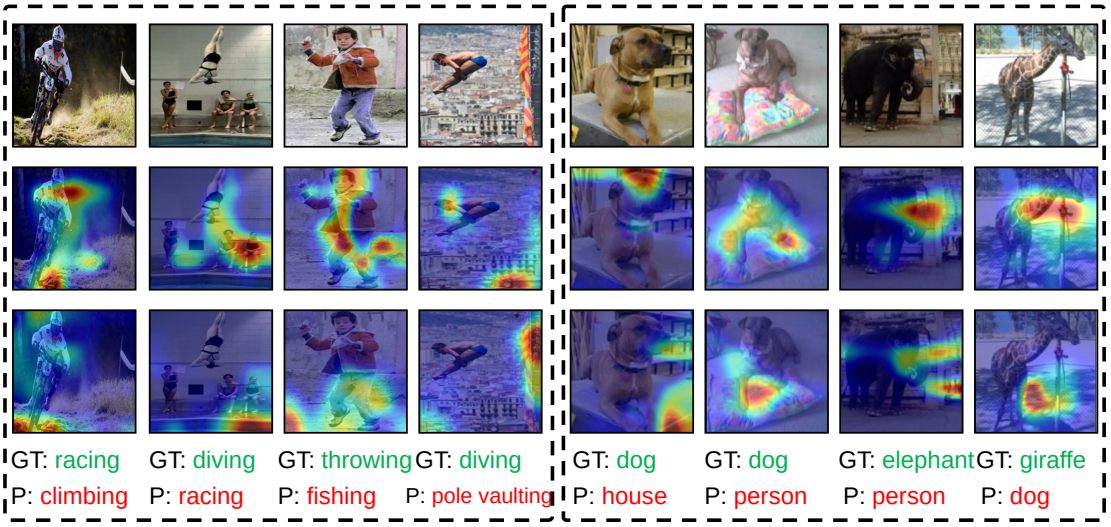


FIGURE 5.12: GradCAM [3] visualizations of IRMCon-IPW failure cases. **Top**: input test images; **Middle**: context visualization by the biased classifier of IRMCon; **Bottom**: class visualization. The four left columns are selected from *BAR* test set, where the model is trained on the 99% biased training set; the four right columns are selected from the *Photo* domain of *PACS*, where the model is trained on the other three domains. GT: ground-truth label; P: predicted label.

ImageNet is collected from the real world and leaks more real images in *Photo*, compare to artificial images in *Cartoon* and *Sketch*. Therefore, we recommend the non-pre-training setting for all debiasing benchmarks to prevent the potential leakage problem.

How to evaluate the context feature learned in IRMCon-IPW? We visualize the comparisons between the context features learned by IRMCon-IPW and LfF in Figure 5.11. We show the training and test performance of the linear classifiers (which are called biased classification heads) that are trained with context features and class labels, *i.e.*, to learn the bias intentionally. As we can see from the figure, ours shows almost the same learning behavior as the upper bound case: the context is invariant to the class and the class should be predicted by random chance. This implies that IRMCon-IPW is able to recover the oracle distribution of contexts in images. This can be taken as a support to the illustration at the bottom in Figure 5.9 where using our weights can achieve a balanced context distribution, the ground truth distribution.

How does IRMCon-IPW tackle domain generalization issues? In contrast to the datasets with pre-defined context distribution in training (*e.g.*, set color

distribution in each class in *Colored MNIST* dataset [2]), the domain generalization dataset such as *PACS* does not have such explicit context settings. But it has implicit context distribution associated with the domain. This distribution is often imbalanced, which leads to the problems of context bias (similar to debiasing datasets, such as *BAR*). Our method can therefore help *PACS* to "debias". We note that our improvement for *PACS* is not as pronounced as that on the debiasing datasets compared to ERM. This may be because the context bias in *PACS* is not as severe as that in debiasing datasets. In addition, the domain generalization dataset encounters another challenge, namely context absence: a context is new in the test set. It is not intuitive how to solve this additional issue by re-weighting methods.

Failure cases. We show some failures of our IRMCon in Figure 5.12. The failure cases are selected if the classification result of IRMCon-IPW is incorrect. As expected, we see that the key reasons for failure are the incorrect context estimations, *i.e.*, contexts are mixed with foreground or incorrectly focus on the foreground. By examining the *BAR* dataset, we find that some contexts, *e.g.*, "pool" for the class "diving", are relatively unique for certain classes. This implies that the contexts are not invariant to class. To resolve this problem, we conjecture that this is a dataset failure and that the only way out is to bring external knowledge.

Chapter 6

Summary

6.1 Conclusion

In this thesis, we point out the reason why the deep models lose their performance in the testing set, *i.e.*, the OOD Generalization problem, which usually has a different data distribution compared to the training, even though they can perfectly fit the training data by achieving small training loss. The reason is the bias, which is induced by spurious correlations between some inputs and output s in the training, such as dialog history and answers in Visual Dialog, and context and class labels in debiasing, while such correlations will not keep in the testing. Learning such bias will greatly degenerate the generalization ability of the deep models. However, the existing methods either ignore the existence of bias or remove bias incompletely, resulting in unsatisfactory performance. To better remove the bias, we categorize OOD Generalization tasks into three camps, *i.e.*, OOD Generalization with Multiple Modalities, with Multiple Domains, and with no Additional Annotations. Then, we conduct specific case studies for each one to analyze the types of bias and design specific algorithms to remove bias for better OOD performance. We also use sufficient experiments to demonstrate the effectiveness of our algorithms in each case. In the following, we summarize our algorithms and contributions for each case in detail.

OOD Generalization with Multiple Modalities. We focused on Visual Dialog and proposed two causal principles to improve VisDial, which are model-agnostic

and can therefore be applied to almost all the existing methods and lead to significant improvements. These principles came from our in-depth causal analysis of the nature of VisDial, however, which has unfortunately been overlooked by the VisDial community. For technical contributions, we provided some implementation examples of how to apply the causal principles to current models. We conducted extensive experiments on the official VisDial dataset and the online evaluation servers, and the promising results demonstrate the effectiveness of our causal principles for removing bias to improve OOD performance.

OOD Generalization with Multiple Domains. We focused on Domain Generalization (DG) and found that learning domain-invariant features by removing domain bias is the key to DG and almost all existing DG methods claim to learn such invariance by imposing a domain-invariant loss. However, we showed that they fail to achieve the invariance because they suffer from another bias caused by spurious invariance, which is invariant in between training domains but not in the testing domains. Therefore, we proposed DOMAIN+ to help them remove the spurious invariance by splitting a new domain consisting of rare samples, which do not contain such spuriously invariant features. The success of choosing these rare samples is due to our proposed cross-domain influence function, which can be estimated efficiently without the need for retraining. We conducted extensive experiments and the results show that DOMAIN+ helps existing methods achieve new SOTAs and outperform the strong ERM baseline, which was not accomplished by current methods.

OOD Generalization with no Additional Annotations. We focused on Debiasing in this camp. First, we proposed a new training framework, TWO, A two-stage debiasing method: Cross-Entropy (CE) biased training followed by Supervised Contrastive-regularized CE unbiased training. The key difference between our algorithm and others is that the former only applies the re-weighting debiasing strategy after the biased CE training, which is used to be considered as a bias selection model and discarded in training. Our finding is that CE training is only partly biased while achieving invariance to the balanced attributes, where such invariance ensures that the subsequent re-weighting does not introduce the "anti-training bias" bias side effect, which makes re-weighting fail. Through our extensive experiments, the proposed algorithm achieves consistent performances across various train/test bias settings.

Second, we proposed a new context estimation algorithm for better re-weighting to remove context bias, namely IRMCon. Since we found that the context imbalance is the main challenge in learning class invariance for Debiasing, and most prior work addresses this challenge by estimating context bias through classifier failures. We showed how they fail and thus proposed our new method called IRM for Context (IRMCon), which disentangles context features directly without context or class supervision. The success of our method is based on the principle, context is invariant to class, which is the neglected side of the common principle in classification, class is invariant to context. Thanks to the class supervision which has been already provided as environments in training data, IRMCon can achieve context invariance by using IRM on the intra-class sample similarity contrastive loss. We used context features for reweighting, *i.e.*, the universal method designed by Inverse Probability Weighting (IPW), and through extensive experiments, we showed that IRMCon-IPW achieves state-of-the-art results on several OOD benchmarks.

Finally, we would like to conclude our core ideas for removing bias for OOD Generalization and make some discussion about whether and how to implement our proposed algorithms. When encountering a new OOD Generalization task, first, we need to analyze the conditions, *i.e.*, whether it has multiple modalities, multiple domains, or no additional annotations are given, and then categorize it into one of the OOD Generalization camps we mentioned in this thesis. If it has multiple modalities, such as Visual Question Answering, we can analyze the causal relationship between each modality and try to find the underlying bias from the causal graph. Although, our proposed causal principles and algorithms for Visual Dialog usually cannot be directly implemented in the new causal graph, due to the different causalities between different multiple modality tasks, some similar approaches, such as cutting the unreasonable links or de-confounding the confounders, as we mentioned in Chapter 3 can help the new task to remove the bias; if it has multiple domains with domain annotations, we can apply the current Domain Generalization method with our DOMAIN+ to directly improve the OOD performance; if there are no additional annotations, which is the most usual case, such as Class Incremental Learning and Few-Shot Learning, we need to analyze the underlying confounders and applying de-confounding method. The de-confounding method can be implemented as re-weighting, and thus our algorithms TWO and IRMCon can be implemented to help remove the confounding bias, and we provide more

discussion for implementing our methods in Class Incremental Learning and Few-Shot Learning in Section 6.2.1. We hope the summarized ideas will enlighten the future for removing the bias of models in other OOD Generalization tasks.

6.2 Future Work

6.2.1 Towards other OOD Generalization tasks

In this thesis, we mainly introduce our proposed algorithms for removing bias in three specific cases in different OOD Generalization camps, and it is easy for us to implement our ideas to design new algorithms for others in the future. We will briefly introduce our design inspirations in the following parts.

Class-Incremental Learning. This OOD Generalization task belongs to OOD Generalization with no additional annotations. Given a full training set for some base classes, this task requires the model to continuously learn new classes without keeping the original data, and most current methods are proposed to mitigate the catastrophic forgetting [45–48]. According to our analysis, we argue that the catastrophic forgetting problem is caused by the bias learning in each step. For example, if we have two base classes, dog and cat and most dogs are brown, the model will use brown to identify dogs, *i.e.*, to learn brown as the class feature, in the ERM (biased) training. If the model finds that there are many brown horses in the coming images, it will realize that brown cannot be the discriminative feature and thus discard it for classification. However, the impact of discarding brown will make the model lose the ability for identifying dogs, which is the underlying reason for the catastrophic forgetting problem.

Therefore, we can apply unbiased training in each step in Class-Incremental Learning to prevent some classes are biased to some class-irrelevant features, just like the context features we have mentioned, and thus the catastrophic forgetting for the learned classes will not happen. Specifically, we can implement our IRMCon-IPW proposed in Section 5.3 to realize unbiased training.

Few-Shot Learning. This OOD Generalization task also belongs to OOD Generalization with no additional annotations. Given a full training set for some base classes, this task requires the model rapidly generalize to new classes with the help

of a small support set, containing several image-label pairs, for these classes [49–53]. Most traditional methods first learn a feature extractor on base classes and then fine-tune a classifier on the support set for classifying the new classes. Based on our analysis, we find that the key challenges are two-fold: 1) training on base classes in the first step will make the feature extractor biased to the class-correlated features, and thus some discriminative features for the new classes but useless for classifying base classes will not be learned, which can not be further learned in the fine-tuning stage where the backbone is frozen; 2) the fine-tuning stage will bring context bias.

To remove the first bias, we will encourage the feature extractor to present more features by some non-classification objectives, such as contrastive loss and reconstruction, and then all features, sometimes we call equivalent features, are extracted for the next fine-tuning stage. Second, we need to conduct unbiased training in fine-tuning. However, different from the previous tasks where we have enough samples for training and we can use re-weighting strategies, re-weighting for such few samples on the small support set will more likely introduce some undesirable bias, just like the bias we have discussed in Section 5.2. Therefore, we propose to use the feature selection strategy, to select the class features based on some causal discovery methods and only fine-tune the classifier on these causal features.

It is worth noting that, recently, fine-tuning a pre-trained model, such as CLIP [173], for Few-Shot Learning become more and more popular. Under this setting, we do not have the first problem because all features we need for classifying any class must be extracted by the full pre-trained backbone, and what we need to do is to conduct feature selection for the support set to remove the bias in the fine-tuning.

6.2.2 Towards improving our algorithm

Besides designing algorithms for other OOD Generalization tasks, we also have some important pending improvements for our algorithm in Section 5.3. As we have discussed, IRMCon is proposed to extract better context features for re-weighting in an intra-class contrastive way. There are some shortcomings: 1) IRMCon may extract a lot of useless context features because any feature that can help to decrease the intra-class contrastive will be learned by IRMCon, but only a part, even a small part of those features will induce the bias, which means

most learned context features will not be used in the weight estimation for re-weighting. 2) the feature learning method, contrastive learning, is out-date, as its ability is limited to data augmentation, batch size, and positive/negative sampling strategies, especially compared to advances in another class-irrelevant objective, reconstruction method, such as MAE and Diffusion Models [174, 175].

Therefore, we design corresponding improvements for IRMCon: First, we train a traditional classification model where the learned features contain class features and biased context features, and then we freeze the backbone of the extractor and use our context learning model to further extract context from the learned features by eliminating class features. In this way, we can only extract the context features causing the bias and do not have to represent every context in the training set. As for replacing the contrastive objective, we can apply a conditional reconstruction strategy by conditioning on class, and then the class features are not necessary to be learned by the context extractor, *i.e.*, class features are removed. Furthermore, we can apply the invariance method to further remove class features by treating classes as environments and minimizing the invariance loss to remove the learning for environments, *i.e.*, classes.

List of Author's Awards and Publications

Awards

- **First place** in the second Visual Dialog Challenge, on CVPR 2019 Visual Question Answering and Visual Dialog Workshop.
  Jiaxin Qi, Yulei Niu, Jianqiang Huang, Hanwang Zhang, Xian-Sheng Hua, and Ji-Rong Wen.

- **Second place** in the third Visual Dialog Challenge, on CVPR 2020 Visual Question Answering Workshop.
  Jiaxin Qi, Tan Wang, Xian-Sheng Hua, and Hanwang Zhang.

- **Jury prize** and 3rd place in the VIPriors Challenge, on ICCV 2021.
  Tan Wang, Wanqi Yin, **Jiaxin Qi**, Jin Liu, Jayashree Karlekar, and Hanwang Zhang.

Conference Proceedings

- **Jiaxin Qi**, Yulei Niu, Jianqiang Huang, and Hanwang Zhang. Two Causal Principles for Improving Visual Dialog. in *CVPR 2020*.

- Jianqiang Huang, Yu Qin, **Jiaxin Qi**, Qianru Sun, and Hanwang Zhang. Deconfounded Visual Grounding. in *AAAI 2022*.

- Kaihua Tang, Mingyuan Tao, **Jiaxin Qi**, Zhenguang Liu, and Hanwang Zhang. Invariant Feature Learning for Generalized Long-Tailed Classification. in *ECCV 2022*.

- **Jiaxin Qi**, Kaihua Tang, Qianru Sun, Xian-Sheng Hua, and Hanwang Zhang. Class Is Invariant to Context and Vice Versa: On Learning Invariance for Out-Of-Distribution Generalization. in *ECCV 2022*.

- **Jiaxin Qi**, Qianru Sun, Xian-Sheng Hua, and Hanwang Zhang. A Two-stage Method for Training Unbiased Models. *Under Review*.

- **Jiaxin Qi***, Zike Wu*, Qianru Sun, Xian-Sheng Hua, and Hanwang Zhang. DOMAIN+: Splitting a New Influential Domain for Domain Generalization.[1] *Under Review*.

[1]The superscript $*$ indicates the equal contributions

Bibliography

[1] Laurens Van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. *Journal of machine learning research*, 9(11), 2008. xviii, xix, 52, 55, 84

[2] Junhyun Nam, Hyuntak Cha, Sungsoo Ahn, Jaeho Lee, and Jinwoo Shin. Learning from failure: De-biasing classifier from biased classifier. *Advances in Neural Information Processing Systems*, 2020. xviii, xix, xx, xxii, 7, 11, 12, 15, 43, 59, 60, 61, 62, 63, 64, 66, 67, 68, 69, 70, 71, 73, 75, 77, 80, 82, 83, 84, 85, 86, 87, 89, 91

[3] Ramprasaath R Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *Proceedings of the IEEE International Conference on Computer Vision*, 2017. xviii, xix, xx, 73, 75, 90

[4] Martin Arjovsky, Léon Bottou, Ishaan Gulrajani, and David Lopez-Paz. Invariant risk minimization. *arXiv preprint arXiv:1907.02893*, 2019. xix, 10, 11, 13, 42, 43, 45, 46, 51, 52, 53, 54, 75, 76, 77, 79, 87

[5] Fabio M Carlucci, Antonio D'Innocente, Silvia Bucci, Barbara Caputo, and Tatiana Tommasi. Domain generalization by solving jigsaw puzzles. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019. 14, 79

[6] Jungsoo Lee, Eungyeup Kim, Juyoung Lee, Jihyeon Lee, and Jaegul Choo. Learning debiased representation via disentangled feature augmentation. In *Advances in Neural Information Processing Systems*, 2021. 12, 15, 44, 60, 63, 66, 67, 68, 69, 70, 71, 75, 80, 84, 85

[7] Ya Li, Xinmei Tian, Mingming Gong, Yajing Liu, Tongliang Liu, Kun Zhang, and Dacheng Tao. Deep domain generalization via conditional invariant adversarial networks. In *Proceedings of the European conference on computer vision (ECCV)*, 2018. 14, 75

[8] Riccardo Volpi and Vittorio Murino. Addressing model vulnerability to distributional shifts over image transformation sets. In *Proceedings of the IEEE International Conference on Computer Vision*, 2019. 14

[9] Xingxuan Zhang, Peng Cui, Renzhe Xu, Linjun Zhou, Yue He, and Zheyan Shen. Deep stable learning for out-of-distribution generalization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021. xix, 15, 74, 77, 84, 87

[10] Visual Dialog Challenge 2019 Leaderboard. `https://evalai.cloudcv.org/web/challenges/challenge-page/161/leaderboard/483/`, 2019. xxi, 17, 39, 40

[11] Da Li, Yongxin Yang, Yi-Zhe Song, and Timothy M Hospedales. Deeper, broader and artier domain generalization. In *Proceedings of the IEEE international conference on computer vision*, 2017. xxii, 1, 7, 51, 60, 83, 87

[12] Ishaan Gulrajani and David Lopez-Paz. In search of lost domain generalization. In *International Conference on Learning Representations*, 2021. xxii, 1, 10, 41, 42, 45, 51, 54, 60, 76, 77, 84, 86, 87, 88, 89

[13] Ian Goodfellow, Yoshua Bengio, and Aaron Courville. *Deep learning*. MIT press, 2016. 1, 60

[14] Robert Geirhos, Patricia Rubisch, Claudio Michaelis, Matthias Bethge, Felix A Wichmann, and Wieland Brendel. Imagenet-trained cnns are biased towards texture; increasing shape bias improves accuracy and robustness. In *International Conference on Learning Representations*, 2019. 1, 8, 41, 60

[15] Zhongqi Yue, Qianru Sun, Xian-Sheng Hua, and Hanwang Zhang. Transporting causal mechanisms for unsupervised domain adaptation. In *Proceedings of the IEEE International Conference on Computer Vision*, 2021. 8, 42

[16] Tan Wang, Chang Zhou, Qianru Sun, and Hanwang Zhang. Causal attention for unbiased visual recognition. In *Proceedings of the IEEE International Conference on Computer Vision*, 2021. 14, 41, 74, 76, 79

[17] Tan Wang, Zhongqi Yue, Jianqiang Huang, Qianru Sun, and Hanwang Zhang. Self-supervised learning disentangled group representation as feature. *Advances in Neural Information Processing Systems*, 34, 2021. 42, 78, 81

[18] Zheyan Shen, Jiashuo Liu, Yue He, Xingxuan Zhang, Renzhe Xu, Han Yu, and Peng Cui. Towards out-of-distribution generalization: A survey. *arXiv preprint arXiv:2108.13624*, 2021. 1, 7, 60

[19] Judea Pearl, Madelyn Glymour, and Nicholas P Jewell. *Causal inference in statistics: A primer*. John Wiley & Sons, 2016. 2, 13, 15, 18, 20, 21, 23, 25, 30, 65, 66

[20] Vladimir Vapnik. Principles of risk minimization for learning theory. In *Advances in neural information processing systems*, 1992. 7, 74

[21] Dan Hendrycks and Thomas Dietterich. Benchmarking neural network robustness to common corruptions and perturbations. In *International Conference on Learning Representations*, 2019. 7

[22] Benjamin Recht, Rebecca Roelofs, Ludwig Schmidt, and Vaishaal Shankar. Do imagenet classifiers generalize to imagenet? In *International Conference on Machine Learning*, 2019.

[23] Dan Hendrycks, Kevin Zhao, Steven Basart, Jacob Steinhardt, and Dawn Song. Natural adversarial examples. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2021. 7

[24] Abhishek Das, Satwik Kottur, Khushi Gupta, Avi Singh, Deshraj Yadav, José MF Moura, Devi Parikh, and Dhruv Batra. Visual dialog. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 326–335, 2017. 7, 9, 17, 18, 20, 26, 34, 35, 36, 37, 39

[25] Yulei Niu, Hanwang Zhang, Manli Zhang, Jianhong Zhang, Zhiwu Lu, and Ji-Rong Wen. Recursive visual attention in visual dialog. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6679–6688, 2019. 20, 27, 34, 35, 36, 37

[26] Tianhao Yang, Zheng-Jun Zha, and Hanwang Zhang. Making history matter: History-advantage sequence training for visual dialog. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2561–2569, 2019. 7, 9, 18, 26

[27] Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C Lawrence Zitnick, and Devi Parikh. Vqa: Visual question answering. In *Proceedings of the IEEE international conference on computer vision*, pages 2425–2433, 2015. 7, 9, 17

[28] Meet Shah, Xinlei Chen, Marcus Rohrbach, and Devi Parikh. Cycle-consistency for robust visual question answering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6649–6658, 2019.

[29] Long Chen, Xin Yan, Jun Xiao, Hanwang Zhang, Shiliang Pu, and Yueting Zhuang. Counterfactual samples synthesizing for robust visual question answering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10800–10809, 2020. 7

[30] Krikamol Muandet, David Balduzzi, and Bernhard Schölkopf. Domain generalization via invariant feature representation. In *International Conference on Machine Learning*, 2013. 7

[31] Ryo Okumura, Masashi Okada, and Tadahiro Taniguchi. Domain-adversarial and-conditional state space model for imitation learning. In *2020 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 2020. 11, 84

[32] Baochen Sun and Kate Saenko. Deep coral: Correlation alignment for deep domain adaptation. In *European conference on computer vision*, pages 443–450. Springer, 2016. 7, 11, 42, 45, 52, 53, 54

[33] Christopher Clark, Mark Yatskar, and Luke Zettlemoyer. Don't take the easy way out: Ensemble based methods for avoiding known dataset biases. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing, EMNLP-IJCNLP 2019, Hong Kong, China, November 3-7, 2019*. Association for Computational Linguistics, 2019. 8

[34] Byungju Kim, Hyunwoo Kim, Kyungsu Kim, Sungjin Kim, and Junmo Kim. Learning not to learn: Training deep neural networks with biased data. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9012–9020, 2019.

[35] Yi Li and Nuno Vasconcelos. Repair: Removing representation bias by dataset resampling. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019.

[36] Haohan Wang, Zexue He, Zachary C Lipton, and Eric P Xing. Learning robust representations by projecting superficial statistics out. In *International Conference on Learning Representations*, 2019. 8

[37] Shai Ben-David, John Blitzer, Koby Crammer, Fernando Pereira, et al. Analysis of representations for domain adaptation. *Advances in Neural Information Processing Systems*, 2007. 8

[38] Mingming Gong, Kun Zhang, Tongliang Liu, Dacheng Tao, Clark Glymour, and Bernhard Schölkopf. Domain adaptation with conditional transferable components. In *International Conference on Machine Learning*, 2016.

[39] Eric Tzeng, Judy Hoffman, Kate Saenko, and Trevor Darrell. Adversarial discriminative domain adaptation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2017.

[40] Judy Hoffman, Eric Tzeng, Taesung Park, Jun-Yan Zhu, Phillip Isola, Kate Saenko, Alexei Efros, and Trevor Darrell. Cycada: Cycle-consistent adversarial domain adaptation. In *International conference on machine learning*. Pmlr, 2018. 8

[41] Salman H Khan, Munawar Hayat, Mohammed Bennamoun, Ferdous A Sohel, and Roberto Togneri. Cost-sensitive learning of deep feature representations from imbalanced data. *IEEE transactions on neural networks and learning systems*, 2017. 8

[42] Ziwei Liu, Zhongqi Miao, Xiaohang Zhan, Jiayun Wang, Boqing Gong, and Stella X Yu. Large-scale long-tailed recognition in an open world. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019.

[43] Kaihua Tang, Jianqiang Huang, and Hanwang Zhang. Long-tailed classification by keeping the good and removing the bad momentum causal effect.

Advances in Advances in Neural Information Processing Systems, 2020. 12, 13

[44] Beier Zhu, Yulei Niu, Xian-Sheng Hua, and Hanwang Zhang. Cross-domain empirical risk minimization for unbiased long-tailed classification. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 2022. 8

[45] Zhizhong Li and Derek Hoiem. Learning without forgetting. *IEEE transactions on pattern analysis and machine intelligence*, 2017. 8, 96

[46] Sylvestre-Alvise Rebuffi, Alexander Kolesnikov, Georg Sperl, and Christoph H Lampert. icarl: Incremental classifier and representation learning. In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, 2017.

[47] Saihui Hou, Xinyu Pan, Chen Change Loy, Zilei Wang, and Dahua Lin. Learning a unified classifier incrementally via rebalancing. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019.

[48] Tz-Ying Wu, Gurumurthy Swaminathan, Zhizhong Li, Avinash Ravichandran, Nuno Vasconcelos, Rahul Bhotika, and Stefano Soatto. Class-incremental learning with strong pre-trained models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022. 8, 96

[49] Jake Snell, Kevin Swersky, and Richard Zemel. Prototypical networks for few-shot learning. *Advances in neural information processing systems*, 2017. 8, 97

[50] Flood Sung, Yongxin Yang, Li Zhang, Tao Xiang, Philip HS Torr, and Timothy M Hospedales. Learning to compare: Relation network for few-shot learning. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018.

[51] Qianru Sun, Yaoyao Liu, Tat-Seng Chua, and Bernt Schiele. Meta-transfer learning for few-shot learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019.

[52] Wei-Yu Chen, Yen-Cheng Liu, Zsolt Kira, Yu-Chiang Frank Wang, and Jia-Bin Huang. A closer look at few-shot classification. *International Conference on Learning Representations*, 2019.

[53] Yaqing Wang, Quanming Yao, James T Kwok, and Lionel M Ni. Generalizing from a few examples: A survey on few-shot learning. *ACM computing surveys (csur)*, 2020. 8, 97

[54] Harm De Vries, Florian Strub, Sarath Chandar, Olivier Pietquin, Hugo Larochelle, and Aaron Courville. Guesswhat?! visual object discovery through multi-modal dialogue. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5503–5512, 2017. 9

[55] Ting Yao, Yingwei Pan, Yehao Li, and Tao Mei. Exploring visual relationship for image captioning. In *Proceedings of the European Conference on Computer Vision*, pages 684–699, 2018. 9

[56] Xu Yang, Kaihua Tang, Hanwang Zhang, and Jianfei Cai. Auto-encoding scene graphs for image captioning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 10685–10694, 2019.

[57] Peter Anderson, Xiaodong He, Chris Buehler, Damien Teney, Mark Johnson, Stephen Gould, and Lei Zhang. Bottom-up and top-down attention for image captioning and visual question answering. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6077–6086, 2018. 9, 35

[58] Damien Teney, Peter Anderson, Xiaodong He, and Anton van den Hengel. Tips and tricks for visual question answering: Learnings from the 2017 challenge. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4223–4232, 2018. 9, 17

[59] Kaihua Tang, Hanwang Zhang, Baoyuan Wu, Wenhan Luo, and Wei Liu. Learning to compose dynamic tree structures for visual contexts. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6619–6628, 2019. 9

[60] Unnat Jain, Svetlana Lazebnik, and Alexander G Schwing. Two can play this game: visual dialog with discriminative question generation and answering. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5754–5763, 2018. 9, 18

[61] Dalu Guo, Chang Xu, and Dacheng Tao. Image-question-answer synergistic network for visual dialog. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 10434–10443, 2019.

[62] Paul Hongsuck Seo, Andreas Lehrmann, Bohyung Han, and Leonid Sigal. Visual reference resolution using attention memory for visual dialog. In *Advances in neural information processing systems*, pages 3719–3729, 2017.

[63] Zhe Gan, Yu Cheng, Ahmed EI Kholy, Linjie Li, Jingjing Liu, and Jianfeng Gao. Multi-step reasoning via recurrent dual attention for visual dialog. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, 2019. 17, 18, 40

[64] Idan Schwartz, Seunghak Yu, Tamir Hazan, and Alexander G Schwing. Factor graph attention. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2039–2048, 2019. 18, 40

[65] Zilong Zheng, Wenguan Wang, Siyuan Qi, and Song-Chun Zhu. Reasoning visual dialogs with structural and partial observations. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6669–6678, 2019. 9

[66] Gi-Cheon Kang, Jaeseo Lim, and Byoung-Tak Zhang. Dual attention networks for visual reference resolution in visual dialog. *arXiv preprint arXiv:1902.09368*, 2019. 9, 18, 40

[67] Qi Wu, Peng Wang, Chunhua Shen, Ian Reid, and Anton van den Hengel. Are you talking to me? reasoned visual dialog generation through adversarial learning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6106–6115, 2018. 9, 18, 20, 26, 27, 34, 35, 36, 37

[68] Satwik Kottur, José MF Moura, Devi Parikh, Dhruv Batra, and Marcus Rohrbach. Visual coreference resolution in visual dialog using neural module networks. In *Proceedings of the European Conference on Computer Vision*, pages 153–169, 2018. 9

[69] Zeyi Huang, Haohan Wang, Eric P Xing, and Dong Huang. Self-challenging improves cross-domain generalization. In *European Conference on Computer Vision*, pages 124–140. Springer, 2020. 10, 52, 53, 84, 87

[70] Da Li, Yongxin Yang, Yi-Zhe Song, and Timothy Hospedales. Learning to generalize: Meta-learning for domain generalization. In *Proceedings of the AAAI conference on artificial intelligence*, 2018. 52, 53, 54, 87

[71] Mohammad Pezeshki, Oumar Kaba, Yoshua Bengio, Aaron C Courville, Doina Precup, and Guillaume Lajoie. Gradient starvation: A learning proclivity in neural networks. *Advances in Neural Information Processing Systems*, 34:1256–1272, 2021. 87

[72] Daehee Kim, Youngjun Yoo, Seunghyun Park, Jinkyu Kim, and Jaekoo Lee. Selfreg: Self-supervised contrastive regularization for domain generalization. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 9619–9628, 2021. 10

[73] David Krueger, Ethan Caballero, Joern-Henrik Jacobsen, Amy Zhang, Jonathan Binas, Dinghuai Zhang, Remi Le Priol, and Aaron Courville. Out-of-distribution generalization via risk extrapolation (rex). In *International Conference on Machine Learning*, pages 5815–5826. PMLR, 2021. 10, 11, 13, 43, 87

[74] Shiori Sagawa, Pang Wei Koh, Tatsunori B Hashimoto, and Percy Liang. Distributionally robust neural networks. In *International Conference on Learning Representations*, 2019. 11, 52, 53, 54, 87

[75] Kartik Ahuja, Ethan Caballero, Dinghuai Zhang, Jean-Christophe Gagnon-Audet, Yoshua Bengio, Ioannis Mitliagkas, and Irina Rish. Invariance principle meets information bottleneck for out-of-distribution generalization. *Advances in Neural Information Processing Systems*, 34:3438–3450, 2021. 10

[76] Haoliang Li, Sinno Jialin Pan, Shiqi Wang, and Alex C Kot. Domain generalization with adversarial feature learning. In *Proceedings of the IEEE/CVF*

Conference on Computer Vision and Pattern Recognition, 2018. 11, 43, 52, 53, 54, 84

[77] Yuge Shi, Jeffrey Seely, Philip HS Torr, N Siddharth, Awni Hannun, Nicolas Usunier, and Gabriel Synnaeve. Gradient matching for domain generalization. In *International Conference on Learning Representations*, 2022. 11, 42, 43, 45, 52, 53, 54, 84, 87

[78] Giambattista Parascandolo, Alexander Neitz, Antonio Orvieto, Luigi Gresele, and Bernhard Schölkopf. Learning explanations that are hard to vary. In *International Conference on Learning Representations*, 2021. 52, 53, 54

[79] Alexandre Rame, Corentin Dancette, and Matthieu Cord. Fishr: Invariant gradient variances for out-of-distribution generalization. In *International Conference on Machine Learning*, pages 18347–18377. PMLR, 2022. 11

[80] Hana Ajakan, Pascal Germain, Hugo Larochelle, François Laviolette, and Mario Marchand. Domain-adversarial neural networks. In *Advances in Neural Information Processing Systems*, 2014. 11, 84, 87

[81] Evan Z Liu, Behzad Haghgoo, Annie S Chen, Aditi Raghunathan, Pang Wei Koh, Shiori Sagawa, Percy Liang, and Chelsea Finn. Just train twice: Improving group robustness without training group information. In *International Conference on Machine Learning*, pages 6781–6792. PMLR, 2021. 12, 43, 67, 69, 70

[82] Hyojin Bahng, Sanghyuk Chun, Sangdoo Yun, Jaegul Choo, and Seong Joon Oh. Learning de-biased representations with biased representations. In *International Conference on Machine Learning*, 2020. 12, 43, 59, 67, 69, 70, 73, 75, 84, 85

[83] Eungyeup Kim, Jihyeon Lee, and Jaegul Choo. Biaswap: Removing dataset bias with bias-tailored swapping augmentation. In *Proceedings of the IEEE International Conference on Computer Vision*, 2021. 12, 60

[84] Youngkyu Hong and Eunho Yang. Unbiased classification through bias-contrastive and bias-balanced learning. *Advances in Neural Information Processing Systems*, 2021. 12, 66, 67, 69, 70

[85] Prannay Khosla, Piotr Teterwak, Chen Wang, Aaron Sarna, Yonglong Tian, Phillip Isola, Aaron Maschinot, Ce Liu, and Dilip Krishnan. Supervised contrastive learning. *Advances in Neural Information Processing Systems*, 2020. 12, 62, 64

[86] Bingyi Kang, Saining Xie, Marcus Rohrbach, Zhicheng Yan, Albert Gordo, Jiashi Feng, and Yannis Kalantidis. Decoupling representation and classifier for long-tailed recognition. In *International Conference on Learning Representations*, 2020. 12

[87] Aditya Krishna Menon, Sadeep Jayasumana, Ankit Singh Rawat, Himanshu Jain, Andreas Veit, and Sanjiv Kumar. Long-tail learning via logit adjustment. In *International Conference on Learning Representations*, 2021. 12

[88] Pang Wei Koh and Percy Liang. Understanding black-box predictions via influence functions. In *International Conference on Machine Learning*, pages 1885–1894. PMLR, 2017. 13, 14, 44, 47, 48, 53

[89] Diviyan Kalainathan, Olivier Goudet, Isabelle Guyon, David Lopez-Paz, and Michèle Sebag. Sam: Structural agnostic model, causal discovery and penalized adversarial learning. *arXiv preprint arXiv:1803.04929*, 2018. 13

[90] Suraj Nair, Yuke Zhu, Silvio Savarese, and Li Fei-Fei. Causal induction from visual observations for goal directed tasks. *arXiv preprint arXiv:1910.01751*, 2019.

[91] Ishita Dasgupta, Jane Wang, Silvia Chiappa, Jovana Mitrovic, Pedro Ortega, David Raposo, Edward Hughes, Peter Battaglia, Matthew Botvinick, and Zeb Kurth-Nelson. Causal reasoning from meta-reinforcement learning. *arXiv preprint arXiv:1901.08162*, 2019.

[92] Yoshua Bengio, Tristan Deleu, Nasim Rahaman, Rosemary Ke, Sébastien Lachapelle, Olexa Bilaniuk, Anirudh Goyal, and Christopher Pal. A meta-transfer objective for learning to disentangle causal mechanisms. *arXiv preprint arXiv:1901.10912*, 2019. 13

[93] Tan Wang, Jianqiang Huang, Hanwang Zhang, and Qianru Sun. Visual commonsense r-cnn. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020. 13

[94] Dong Zhang, Hanwang Zhang, Jinhui Tang, Xian-Sheng Hua, and Qianru Sun. Causal intervention for weakly-supervised semantic segmentation. *Advances in Neural Information Processing Systems*, 33:655–666, 2020. 13, 42

[95] Zhongqi Yue, Hanwang Zhang, Qianru Sun, and Xian-Sheng Hua. Interventional few-shot learning. *Advances in neural information processing systems*, 2020. 13

[96] Kaihua Tang, Yulei Niu, Jianqiang Huang, Jiaxin Shi, and Hanwang Zhang. Unbiased scene graph generation from biased training. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 3716–3725, 2020. 13, 42

[97] Yulei Niu, Kaihua Tang, Hanwang Zhang, Zhiwu Lu, Xian-Sheng Hua, and Ji-Rong Wen. Counterfactual vqa: A cause-effect look at language bias. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021. 13, 42

[98] R. Dennis Cook and Sanford Weisberg. *Residuals and Influence in Regression*. New York: Chapman and Hall, 1982. 13, 47

[99] Samyadeep Basu, Xuchen You, and Soheil Feizi. On second-order group influence functions for black-box predictions. In *Proceedings of the 37th International Conference on Machine Learning*, pages 715–724. PMLR, November 2020. 13, 44, 56

[100] Jakub Sliwinski, Martin Strobel, and Yair Zick. Axiomatic characterization of data-driven influence measures for classification. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 718–725, 2019. 13

[101] Shuming Kong, Yanyan Shen, and Linpeng Huang. Resolving training biases via influence-based data relabeling. In *International Conference on Learning Representations*, 2022. 14, 56

[102] Zifeng Wang, Hong Zhu, Zhenhua Dong, Xiuqiang He, and Shao-Lun Huang. Less is better: Unweighted data subsampling via influence function. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 6340–6347, 2020. 13

[103] Shen Yan, Huan Song, Nanxiang Li, Lincan Zou, and Liu Ren. Improve unsupervised domain adaptation with mixup training. *arXiv preprint arXiv:2001.00677*, 2020. 14, 52, 53, 54, 87

[104] Divyat Mahajan, Shruti Tople, and Amit Sharma. Domain generalization using causal matching. In *International Conference on Machine Learning*, pages 7313–7324. PMLR, 2021. 14

[105] Jonas Peters, Peter Bühlmann, and Nicolai Meinshausen. Causal inference by using invariant prediction: identification and confidence intervals. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 78(5): 947–1012, 2016. 42

[106] Niklas Pfister, Peter Bühlmann, and Jonas Peters. Invariant causal prediction for sequential data. *Journal of the American Statistical Association*, 114(527): 1264–1276, 2019. 14

[107] Paul R Rosenbaum and Donald B Rubin. The central role of the propensity score in observational studies for causal effects. *Biometrika*, 1983. 15, 25, 66

[108] Donald B Rubin. Estimating causal effects from large data sets using propensity scores. *Annals of internal medicine*, 1997. 15, 25, 66

[109] Zhilu Zhang and Mert Sabuncu. Generalized cross entropy loss for training deep neural networks with noisy labels. *Advances in neural information processing systems*, 31, 2018. 15, 43, 80

[110] Jiasen Lu, Jianwei Yang, Dhruv Batra, and Devi Parikh. Hierarchical question-image co-attention for visual question answering. In *Advances In Neural Information Processing Systems*, pages 289–297, 2016. 17

[111] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in neural information processing systems*, pages 5998–6008, 2017. 17

[112] Visual Dialog. `https://visualdialog.org/`, 2019. 17, 26, 34, 35

[113] Zhou Yu, Jun Yu, Jianping Fan, and Dacheng Tao. Multi-modal factorized bilinear pooling with co-attention learning for visual question answering. In *Proceedings of the IEEE international conference on computer vision*, pages 1821–1830, 2017. 17

[114] Hyeonseob Nam, Jung-Woo Ha, and Jeonghee Kim. Dual attention networks for multimodal reasoning and matching. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 299–307, 2017. 17

[115] Judea Pearl et al. Causal inference in statistics: An overview. *Statistics surveys*, 3:96–146, 2009. 18

[116] Jiasen Lu, Anitha Kannan, Jianwei Yang, Devi Parikh, and Dhruv Batra. Best of both worlds: Transferring knowledge from discriminative learning to a generative visual dialog model. In *Advances in Neural Information Processing Systems*, pages 314–324, 2017. 20, 26, 34, 35, 36, 37

[117] Yonghan Jung, Jin Tian, and Elias Bareinboim. Learning causal effects via weighted empirical risk minimization. *Advances in Neural Information Processing Systems*, 2020. 25, 79

[118] Shaun R Seaman and Stijn Vansteelandt. Introduction to double robust methods for incomplete data. *Statistical science: a review journal of the Institute of Mathematical Statistics*, 2018. 25, 79

[119] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. In *Advances in neural information processing systems*, pages 91–99, 2015. 26

[120] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2016. 26, 74

[121] Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, Li-Jia Li, David A Shamma, et al. Visual genome: Connecting language and vision using crowdsourced dense image annotations. *International Journal of Computer Vision*, 123(1): 32–73, 2017. 26

[122] Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural computation*, 9(8):1735–1780, 1997. 26

[123] Kyunghyun Cho, Bart van Merrienboer, Dzmitry Bahdanau, and Yoshua Bengio. On the properties of neural machine translation: Encoder-decoder approaches. In *Proceedings of SSST@EMNLP 2014, Eighth Workshop on Syntax, Semantics and Structure in Statistical Translation, Doha, Qatar, 25 October 2014*. Association for Computational Linguistics, 2014. 26

[124] Kelvin Xu, Jimmy Ba, Ryan Kiros, Kyunghyun Cho, Aaron Courville, Ruslan Salakhudinov, Rich Zemel, and Yoshua Bengio. Show, attend and tell: Neural image caption generation with visual attention. In *International conference on machine learning*, pages 2048–2057, 2015. 26, 33

[125] Judea Pearl and Dana Mackenzie. *THE BOOK OF WHY: THE NEW SCIENCE OF CAUSE AND EFFECT*. Basic Books, 2018. 30

[126] Zhe Cao, Tao Qin, Tie-Yan Liu, Ming-Feng Tsai, and Hang Li. Learning to rank: from pairwise approach to listwise approach. In *Proceedings of the 24th international conference on Machine learning*, pages 129–136. ACM, 2007. 33

[127] Nitish Srivastava, Geoffrey Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. Dropout: a simple way to prevent neural networks from overfitting. *The journal of machine learning research*, 15(1):1929–1958, 2014. 33, 36

[128] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *European conference on computer vision*, pages 740–755. Springer, 2014. 34, 74

[129] Jeffrey Pennington, Richard Socher, and Christopher Manning. Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pages 1532–1543, 2014. 35

[130] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014. 36, 53, 70, 85

[131] Kaiyang Zhou, Ziwei Liu, Yu Qiao, Tao Xiang, and Chen Change Loy. Domain generalization: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pages 1–20, 2022. ISSN 0162-8828, 2160-9292, 1939-3539. doi: 10.1109/TPAMI.2022.3195549. 41

[132] Jindong Wang, Cuiling Lan, Chang Liu, Yidong Ouyang, Tao Qin, Wang Lu, Yiqiang Chen, Wenjun Zeng, and Philip Yu. Generalizing to unseen domains: A survey on domain generalization. *IEEE Transactions on Knowledge and Data Engineering*, pages 1–1, 2022. ISSN 1558-2191. doi: 10.1109/TKDE.2022.3178128. 41

[133] Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. Focal loss for dense object detection. In *Proceedings of the IEEE international conference on computer vision*, 2017. 43, 48

[134] Jiaxin Qi, Kaihua Tang, Qianru Sun, Xian-Sheng Hua, and Hanwang Zhang. Class is invariant to context and vice versa: on learning invariance for out-of-distribution generalization. In *European Conference on Computer Vision*, pages 92–109. Springer, 2022. 44

[135] Abhinav Shrivastava, Abhinav Gupta, and Ross Girshick. Training region-based object detectors with online hard example mining. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 761–769, 2016. 48

[136] Chen Fang, Ye Xu, and Daniel N Rockmore. Unbiased metric learning: On the utilization of multiple datasets and web images for softening bias. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 1657–1664, 2013. 51

[137] Hemanth Venkateswara, Jose Eusebio, Shayok Chakraborty, and Sethuraman Panchanathan. Deep hashing network for unsupervised domain adaptation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5018–5027, 2017. 51

[138] Sara Beery, Grant Van Horn, and Pietro Perona. Recognition in terra incognita. In *Proceedings of the European conference on computer vision*, pages 456–473, 2018. 51

[139] Vladimir Vapnik. *The Nature of Statistical Learning Theory*. Springer Science & Business Media, November 1999. ISBN 978-0-387-98780-4. 52, 53, 54

[140] Hyeonseob Nam, HyunJae Lee, Jongchan Park, Wonjun Yoon, and Donggeun Yoo. Reducing domain gap by reducing style bias. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8690–8699, 2021. 52, 53, 54

[141] Naman Agarwal, Brian Bullins, and Elad Hazan. Second-order stochastic optimization for machine learning in linear time. *The Journal of Machine Learning Research*, 18(1):4148–4187, 2017. 53

[142] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016. 53

[143] Brian Hu Zhang, Blake Lemoine, and Margaret Mitchell. Mitigating unwanted biases with adversarial learning. In *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society*, 2018. 60

[144] Gabriel Grand and Yonatan Belinkov. Adversarial regularization for visual question answering: Strengths, shortcomings, and side effects. *arXiv preprint arXiv:1906.08430*, 2019. 60

[145] Remi Cadene, Corentin Dancette, Matthieu Cord, Devi Parikh, et al. Rubi: Reducing unimodal biases for visual question answering. *Advances in neural information processing systems*, 2019. 62

[146] Enzo Tartaglione, Carlo Alberto Barbano, and Marco Grangetto. End: Entangling and disentangling deep representations for bias correction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021. 62, 84, 85

[147] Bernhard Schölkopf. Causality for machine learning. *arXiv preprint arXiv:1911.10500*, 2019. 65

[148] Giambattista Parascandolo, Niki Kilbertus, Mateo Rojas-Carulla, and Bernhard Schölkopf. Learning independent causal mechanisms. In *International Conference on Machine Learning*. PMLR, 2018.

[149] Raphael Suter, Djordje Miladinovic, Bernhard Schölkopf, and Stefan Bauer. Robustly disentangled causal mechanisms: Validating deep representations for interventional robustness. In *International Conference on Machine Learning*, 2019. 78

[150] Donald B Rubin. Essential concepts of causal inference: a remarkable history and an intriguing future. *Biostatistics & Epidemiology*, 2019.

[151] Elias Bareinboim and Judea Pearl. Controlling selection bias in causal inference. In *Artificial Intelligence and Statistics*, 2012.

[152] Silvia Chiappa. Path-specific counterfactual fairness. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 2019. 65

[153] Kevin P Murphy. *Machine learning: a probabilistic perspective*. MIT press, 2012. 65

[154] Michel Besserve, Arash Mehrjou, Rémy Sun, and Bernhard Schölkopf. Counterfactuals uncover the modular structure of deep generative models. In *International Conference on Machine Learning*, 2020. 67

[155] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *CVPR*, 2009. 74, 89

[156] Bernhard Schölkopf, Francesco Locatello, Stefan Bauer, Nan Rosemary Ke, Nal Kalchbrenner, Anirudh Goyal, and Yoshua Bengio. Toward causal representation learning. *Proceedings of the IEEE*, 109(5):612–634, 2021. 74

[157] Yue He, Zheyan Shen, and Peng Cui. Towards non-iid image classification: A dataset and baselines. *Pattern Recognition*, 110:107383, 2021. 74

[158] Dan Hendrycks and Kevin Gimpel. A baseline for detecting misclassified and out-of-distribution examples in neural networks. In *International Conference on Learning Representations*, 2017.

[159] Shiyu Liang, Yixuan Li, and Rayadurgam Srikant. Enhancing the reliability of out-of-distribution image detection in neural networks. In *International Conference on Learning Representations*, 2018. 74

[160] Roderick JA Little and Donald B Rubin. *Statistical analysis with missing data*, volume 793. John Wiley & Sons, 2019. 76

[161] Irina Higgins, David Amos, David Pfau, Sebastien Racaniere, Loic Matthey, Danilo Rezende, and Alexander Lerchner. Towards a definition of disentangled representations. *arXiv preprint arXiv:1812.02230*, 2018. 78

[162] Riccardo Volpi, Hongseok Namkoong, Ozan Sener, John Duchi, Vittorio Murino, and Silvio Savarese. Generalizing to unseen domains via adversarial data augmentation. In *Advances in Neural Information Processing Systems*, 2018. 79

[163] Peter C Austin. An introduction to propensity score methods for reducing the effects of confounding in observational studies. *Multivariate behavioral research*, 2011. 79

[164] Jiashuo Liu, Zheyuan Hu, Peng Cui, Bo Li, and Zheyan Shen. Heterogeneous risk minimization. In *International Conference on Machine Learning*, 2021. 79

[165] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *International conference on machine learning*, pages 1597–1607. PMLR, 2020. 81

[166] Jean-Bastien Grill, Florian Strub, Florent Altché, Corentin Tallec, Pierre Richemond, Elena Buchatskaya, Carl Doersch, Bernardo Avila Pires, Zhaohan Guo, Mohammad Gheshlaghi Azar, et al. Bootstrap your own latent-a new approach to self-supervised learning. *Advances in Neural Information Processing Systems*, 33:21271–21284, 2020.

[167] Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. Momentum contrast for unsupervised visual representation learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020. 81

[168] Aaron van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*, 2018. 81

[169] Yilun Xu and Tommi Jaakkola. Learning representations that support robust transfer of predictors. *arXiv preprint arXiv:2110.09940*, 2021. 84, 87

[170] Ekin D Cubuk, Barret Zoph, Jonathon Shlens, and Quoc V Le. Randaugment: Practical automated data augmentation with a reduced search space. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pages 702–703, 2020. 88

[171] Catherine Wah, Steve Branson, Peter Welinder, Pietro Perona, and Serge Belongie. The caltech-ucsd birds-200-2011 dataset. *California Institute of Technology*, 2011. 89

[172] Yongqin Xian, Christoph H Lampert, Bernt Schiele, and Zeynep Akata. Zero-shot learning—a comprehensive evaluation of the good, the bad and the ugly. *IEEE transactions on pattern analysis and machine intelligence*, 2018. 89

[173] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning*. PMLR, 2021. 97

[174] Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked autoencoders are scalable vision learners. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022. 98

[175] Prafulla Dhariwal and Alexander Nichol. Diffusion models beat gans on image synthesis. *Advances in Neural Information Processing Systems*, 2021. 98