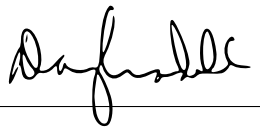
# OUT OF DISTRIBUTION GENERALIZATION, SPURIOUS CORRELATIONS AND SUPERVISED CONTRASTIVE GROUPWISE LEARNING

Jaydeep Chauhan

Submitted to the faculty of the University Graduate School
in partial fulfillment of the requirements
for the degree
Master of Science
in the Department of Computer Science,
Indiana University

May 2023

Accepted by the Graduate Faculty, Indiana University, in partial fulfillment of the requirements for the degree of Master of Science.



Master's Thesis Committee

David J. Crandall, Ph.D.

Hasan Kurban, Ph.D.

defense date

# DEDICATION

*To my parents, Thakor and Anita, for their endless love and support and for giving me the freedom to shape my life.*

# ACKNOWLEDGMENTS

I would like to express my sincere gratitude to my supervisor Dr. David Crandall for his invaluable guidance, support, and encouragement throughout the entire research journey. Honestly, I would not have reached this level without his encouragement and trust in my research ideas.

I would also like to extend my heartfelt thanks to Dr. Hasan Kurban for his time, effort, and feedback in reviewing my thesis and providing valuable insights and suggestions.

Furthermore, I am grateful to my colleagues and friends, specifically Naveksha for providing insightful discussions that help me to improve my research work throughout this research journey.

Last but not least, I would like to express my deep appreciation to my family for their unwavering love, encouragement and support throughout my academic pursuits. Their sacrifices and belief in me have been the driving force behind my success, and I am forever grateful.

Jaydeep Chauhan

# OUT OF DISTRIBUTION GENERALIZATION, SPURIOUS CORRELATIONS AND SUPERVISED CONTRASTIVE GROUPWISE LEARNING

In recent times, deep learning models have exhibited remarkable achievements across a broad range of applications, including image recognition and natural language processing. However, recent research has indicated that these models often encounter difficulties with out-of-distribution (OOD) samples, which vary considerably from the training distributions. Effective OOD generalization is vital for real-world applications. Therefore, this thesis aims to explore the current OOD generalization methods and proposes novel strategies to enhance it.

To begin with, a comprehensive analysis of the existing OOD generalization methods is conducted, and a new benchmark is proposed to compare these methods fairly, using the Waterbirds and CelebA datasets. Moreover, an explainable AI method called 'GradCAM' is employed to visualize the learned feature representation of a model. These visualizations demonstrate that some of the methods indeed facilitate the model to learn generalizable features.

Later in this thesis, a novel group-based reweighted sampling strategy is proposed to improve the feature representation of existing deep learning methods. Additionally, a novel method called 'Supervised Contrastive Groupwise Learning' is proposed to learn a representation that is invariant to spurious attributes in the data. The results show that this approach enhances the generalization capabilities of deep learning models, and the outcomes are comparable to the existing OOD generalization methods in our benchmark.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

# Chapter 1

# Introduction

## 1.1 Brief overview of OOD generalization

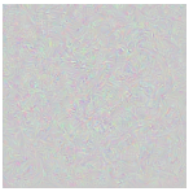
Machine learning methods have achieved tremendous success in solving problems across various fields, such as computer vision (e.g., image classification (Krizhevsky et al. (2012); Simonyan and Zisserman (2015); He et al. (2015); Huang et al. (2018)), object detection (Girshick (2015); Ren et al. (2015); Redmon et al. (2016); He et al. (2017)), image segmentation (Ronneberger et al. (2015); Jegou et al. (2017); Chen et al. (2021), Kirillov et al. (2023))), natural language processing (Mikolov et al. (2013); Bahdanau et al. (2016); Devlin et al. (2019); Radford et al. (2018); Liu et al. (2019)), reinforcement learning (Mnih et al. (2013); Silver et al. (2017)), among others. Much of this progress has been attributed to recent advances in deep learning methods (Lecun et al. (2015)), which involve training large neural network-style architectures to learn hierarchical representations. Generalization beyond the training distribution is one of the long standing goals of machine learning (Lake et al. (2016)). However, to provide any generalization guarantees, a strong and fundamental assumption must be made when developing and training these models: the training and testing data samples are identically and independently distributed (i.e., the i.i.d. assumption (Vapnik (1998))).

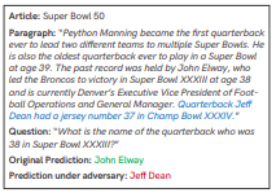
However, when deploying these models in real-world settings, the i.i.d. assumption rarely holds due to distribution shifts and changing environmental conditions, such as variations in background, viewpoint, lighting, or shape in computer vision domain. As a result of the violation of the i.i.d assumption, the model may perform poorly on out-of-distribution datasets, meaning data that differs significantly from the training dataset. Achieving out-

of-distribution generalization is particularly crucial for high-stakes real-world applications, such as autonomous driving, military operations, and healthcare, where the consequences of incorrect predictions can be severe.

### 1.2 Why deep learning models struggle against distribution shifts?

In contemporary deep learning research, numerous research groups are endeavoring to comprehend the challenges posed by out-of-distribution generalization, particularly for deep learning methodologies. Several hypotheses have been proposed to explain this phenomenon. One conjecture asserts that this difficulty may arise from the method we use to train deep learning models. Typically, deep learning models are trained using a widely adopted learning paradigm called 'Empirical Risk Minimization'(ERM). The primary objective of ERM (Vapnik (1992)) is to minimize the average loss across the training dataset. However, training a model using ERM presupposes that the training and testing datasets are identically and independently distributed according to the data generating distribution. Moreover, the generalization bound of ERM expects the ratio between the capacity of the hypothesis and the number of training samples to approach zero. Unfortunately, this ideal circumstance can only be achieved when infinite samples are available during train-






| | | | | |
|---|---|---|---|---|
| **Task for DNN** | Caption image | Recognise object | Recognise pneumonia | Answer question |
| **Problem** | Describes green hillside as grazing sheep | Hallucinates teapot if certain patterns are present | Fails on scans from new hospitals | Changes answer if irrelevant information is added |
| **Shortcut** | Uses background to recognise primary object | Uses features irrecognisable to humans | Looks at hospital token, not lung | Only looks at last sentence and ignores context |

Figure 1.1: Examples of shortcut learning in deep learning models for some real world applications (adapted from Geirhos et al. (2021)).

(A) **Cow: 0.99**, Pasture: 0.99, Grass: 0.99, No Person: 0.98, Mammal: 0.98

(B) No Person: 0.99, Water: 0.98, Beach: 0.97, Outdoors: 0.97, Seashore: 0.97

(C) No Person: 0.97, **Mammal: 0.96**, Water: 0.94, Beach: 0.94, Two: 0.94

Figure 1.2: Image classification models fails to a distribution shifts. (A) Cows are spuriously correlated with greener backgrounds and classified correctly (B) Classified poorly while background changes (C) Top five labels and confidence (adapted from Beery et al. (2018)).

ing, which is impractical in the real world. Any violation of these assumptions during the training process may cause the model to absorb spurious correlations or shortcuts that are not valid under distributional shifts, in order to achieve zero test error. As a result, the model exhibits undesirable behavior during inference on data samples that do not follow the training distribution. Rather than learning invariant and causal relationships, the model relies on spurious correlations that fail to generalize under distributional shifts.

Many prior studies (Rosenfeld et al. (2018); Beery et al. (2018); Geirhos et al. (2021); Dietterich (2019); Geirhos et al. (2022); Ghosal et al. (2022)) have demonstrated the impact of spurious correlation and shortcut learning on the out-of-distribution generalization problem for deep learning models. For example, the well-known cow-camel classification problem in the out-of-domain generalization literature (Beery et al. (2018)) illustrates this issue. In the training distribution, the label is spuriously correlated with the background: most cows are typically found against a greener background, while camels are often against a sandy one. During test time, if this correlation between label and background completely flips (e.g., cows in a desert and camels in a greener background), the accuracy on this new

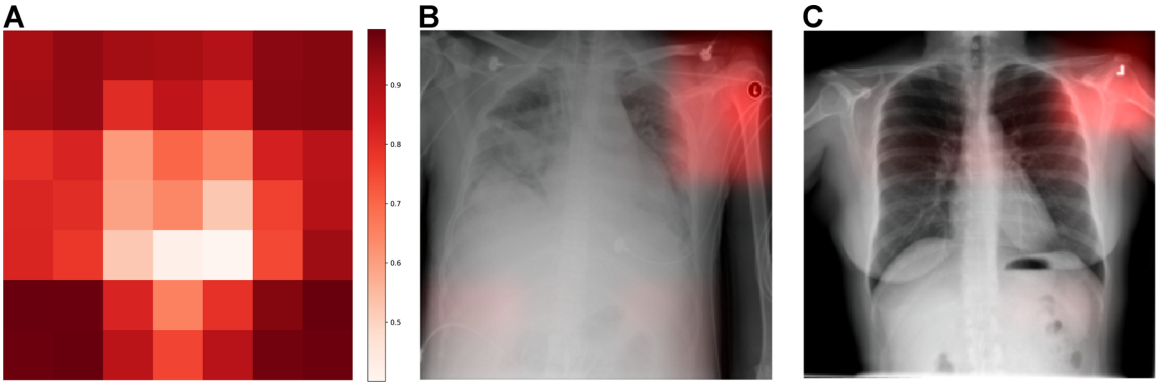



Figure 1.3: CNNs are trained to detect pneumonia relies on spurious features(hospital tokens). (A) Activation heatmaps are average over a sample of images to understand which regions are responsible for trigerring classifier's decision and strong response is coming from the corners. (B-C) shows CNN heavily relies on hospital tokens that are placed in the corner of image to predict pneumonia (adapted from Zech et al. (2018)).

test set drops drastically compared to the training dataset. This phenomenon is also known as shortcut learning (Geirhos et al. (2021)), and prior research has demonstrated its occurrence in a broader range of problems. Another concerning example (Zech et al. (2018)) is the use of CNNs to detect pneumonia, as it has been shown that they rely on non-semantic hospital-specific tokens that radiology technicians places on the patient in the corner of the image while capturing chest X-ray. Additionally, some research (Buolamwini and Gebru (2018); Hashimoto et al. (2018); Rahmattalabi et al. (2020); Tatman (2017)) highlighted how spurious correlations can disproportionately harm minority groups in the data.

There are additional hypotheses (Nakkiran et al. (2019); Sagawa et al. (2020)) which suggest that overparameterization in neural networks may pose a subtle risk for out-of-distribution generalization. In this overparameterized regime, deep learning models may resort to memorizing spurious correlations instead of learning causal features, resulting in reduced performance under distribution shift, particularly for minority groups. Another hypothesis (Nagarajan et al. (2021); Shah et al. (2020); Rahaman et al. (2019); Smith et al. (2021)) proposes that stochastic gradient descent (SGD) has an implicit regularization bias towards learning simpler (low capacity) solutions. Unfortunately, detecting biases and spu-

rious correlations is often easier than identifying task-specific features. For instance, in the camel-cow classification task, it may be easier to learn to detect the greener/sandy background than to distinguish between cows and camels in images. Although the inductive bias of SGD is useful in combating overfitting, it can sometimes force the model to incorporate spurious correlations, which can be detrimental to out-of-distribution samples.

### 1.3 Goals and Motivations

This thesis is motivated by some critical challenges in existing out-of-distribution generalization domain. Despite recent advances in the out-of-distribution generalization literature, there remain significant limitations.

Firstly, one of the open problems is how to formally specify distribution shifts. Since training and testing samples can come from different distributions, and there are numerous ways in which distribution shift can occur in the real world, various research directions are attempting to solve the problem of out-of-distribution generalization, and these methods focus on different ways of modeling distribution shifts.

Secondly, there are no generic benchmarking datasets to keep track of progress in this domain, leading to the generation of numerous datasets and evaluation metrics. Consequently, comparing and evaluating different methods is incredibly challenging.

Thirdly, there is no oracle method to solve the out-of-distribution generalization problem. Thus, without making some assumptions about distribution shifts, it is impossible to solve the problem. Different research work makes different assumptions, such as that training data is pooled from various training environments with environmental labels available during training, or that training data is decomposed into different subsets (groups) with group labels available during training.

Recently, contrastive learning-based methods have gained popularity in the semi-supervised representation learning domain, and they have shown tremendous success for various downstream tasks in computer vision. These methods can learn invariant representations and improve the generalization of deep learning models. However, very little work has attempted to apply these methods to the out-of-distribution generalization domain.

The goal of this thesis is to examine the group-based reweighted sampling strategy and propose a supervised contrastive groupwise learning framework for image classification problems in computer vision. It aims to emphasize the importance of prior knowledge of group labels in addressing spurious correlations and understanding its impact on out-of-distribution generalization in deep learning models. Additionally, the thesis intends to outline a set of future research directions in this field.

### 1.4 Contributions

This thesis includes the following contributions:

1. A detailed survey of the existing methods for out of distribution generalization.
2. Creation of an evaluation benchmark to enable a fair comparison of different out of distribution generalization methods.
3. Utilization of XAI methods, such as Grad-CAM, to visually evaluate the extent to which these methods help in learning task specific robust features.
4. Proposal of a group-based reweighted sampling strategy to investigate the generalization of empirical risk minimization (ERM) and other out of distribution generalization methods.
5. Development of a novel supervised contrastive groupwise learning method for out of distribution generalization.

# Chapter 2

## Background and related work

This chapter will introduce the problem background, and existing research directions to address the out of distribution generalization. This chapter is organized as follows: Section 2.1 introduces the mathematical overview of out of distribution generalization problem. Section 2.2 discusses the categorization and overview of distribution shifts. Section 3.3 provides brief overview of current OOD methods and afterwards from Section 3.4 to Section 3.6 reviews some of the popular OOD methods in the literature that are used in benchmarking. Section 3.7 focuses on Supervised contrastive learning methods. Finally, section 3.8 discusses an Explainable AI method, specifically 'GradCAM'.

### 2.1 Problem definition of OOD generalization

In the context of a supervised learning framework, input vector $\mathcal{X}$ and corresponding labels $\mathcal{Y}$ are represented. A learnable parameter $\Theta$ is used to map the input vector $\mathcal{X}$ to its corresponding label $\mathcal{Y}$ via the machine learning model, $f_\theta : \mathcal{X} \rightarrow \mathcal{Y}$. The discrepancy between the ground truth and predicted labels is measured using a loss function, $L : \mathcal{Y} \times \mathcal{Y} \rightarrow \mathbb{R}$.

To construct a training dataset, $\{(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)\}$, n sample pairs of $\mathcal{X}$ and $\mathcal{Y}$ are drawn from the underlying ground truth training distribution $P_{tr}(\mathcal{X}, \mathcal{Y})$. The goal of machine learning is to develop a model that generalizes well on the unseen test set during training. The samples from the test set are also drawn from the test distribution $P_{tst}(\mathcal{X}, \mathcal{Y})$, where both the training and test distributions are assumed to be drawn from the same underlying ground truth distribution under the i.i.d assumption.

Using the i.i.d assumption, the model is trained with Empirical Risk Minimization (ERM), which minimizes the average training loss of a batch of samples during training to find the optimal parameters of the model using any gradient-based optimization algorithm,

$$L_{\text{ERM}} = \frac{1}{n} \sum_{i=1}^{n} L(f_\theta(x_i), y_i). \tag{2.1}$$

However, in the real world, it is unlikely that the data distribution encountered during deployment will be the same as the training distribution, due to nonstationarity and data distribution shifts. In production, test samples are drawn from a new but slightly shifted distribution $\mathrm{P}'_{\text{tst}}(\mathcal{X}, \mathcal{Y})$. The goal of OOD generalization is to develop a model that is not only optimal for the original test distribution $P_{tst}$ but also for the new shifted distribution $P'_{tst}$.

Designing a machine learning model that can maintain robustness across any type of distribution shift without making any underlying assumptions about the new distribution $P'_{tst}$ remains a formidable challenge. Given the multifaceted nature of the real world, which exhibits an infinite variety of distribution shifts, it is crucial to categorize these shifts and introduce appropriate assumptions in order to address the problem in a more feasible manner.

## 2.2 Categorization of distribution shifts

Distribution shifts can be categorized into three types: Concept shift, Covariate shift, and label shift.

To understand these distribution shifts, we must revisit our problem definition. The goal of supervised machine learning is to estimate a function $f_\theta$ that maps input ($\mathcal{X}$) to target label ($\mathcal{Y}$) from the training dataset. Thus, optimizing the training objective of ERM,

the model approximates the conditional probability distribution between the input vector and label represented as $P(\mathcal{Y}|\mathcal{X})$.

Covariate shift can be defined as a change in the marginal distribution of the input data/features $P(\mathcal{X})$, while the functional relationship between $\mathcal{Y}$ and $\mathcal{X}$ remains the same, in other words $P(\mathcal{Y}|\mathcal{X})$ stays constant. Covariate shift is prevalent in the real world. For example, in an image classification problem, during testing, the lighting, background, and pose may change, but these changes should not affect the relationship between the images and labels. Most methods in the OOD literature are focused on these types of distribution shifts. Covariate shift is closely related to spurious correlations and causality. Since the conditional probability distribution between inputs and labels remains the same and is invariant across different distributions, prior work on causality and invariant representation learning-based methods aims to capture these invariant features that generalize across different distributions.

Concept shift can be defined as a change in the conditional distribution between inputs and labels across different distributions. For some machine learning applications, after deployment, this conditional distribution starts changing over time. For example, in product recommendation, consumer preferences change over time, introducing concept shift. Another example is spam classification, where the types of emails people send change over time. A new type of spam email that the model has never seen before or a change in the words or phrases people use in their emails can cause the conditional probability distribution between labels and features to change. Active learning-based methods address this problem of concept shift.

Label shift is the opposite of covariate shift. In label shift, the marginal distribution of the label's changes over time. Label shift occurs when the class distribution changes between the development and deployment of the model in the wild setting. The class im-

balance problem is one realization of label shift. Methods such as subsampling, importance reweighted sampling, and active learning-based methods are trying to address this problem.

In this thesis, we will focus on covariate shift, and in the following section, we will review some recent work that has tried to address this problem.

### 2.3 Brief over view of OOD methods

OOD methods (Dherin et al. (2021)) can be further divided into three categories: unsupervised representation learning methods, supervised representation learning, and optimization-based methods.

Unsupervised representation learning methods can further be divided into two subcategories: disentangled representation learning and causal learning. Disentangled representation learning ((Bengio et al. (2014); Higgins et al. (2017); Mnih and Kim (2018); Leeb et al. (2021); Dittadi et al. (2021); Creager et al. (2021))) focuses on learning representations which can be factored and disentangled. It is believed that disentangled representation can aid in OOD generalization by separating out factors of variations (spurious and causal factors). For example, in the cow-camel classification task, flipping a background during test time would result in a very sparse change in model representation as it will only affect the spurious factors, and other parts of the representation will be intact. Causal learning-based methods (Locatello et al. (2021); Shen et al. (2021); Yang et al. (2021)) focus on learning a causal graph of the latent data generation process in a semi-supervised or unsupervised way. These approaches are based on combining variational inference and structural causal methods.

Supervised representation learning methods can be divided into two subcategories: domain generalization-based methods and causal invariant representation learning. Domain generalization-based methods incorporate data from different domains to learn a model

that can generalize well to some unseen domain. The goal of domain generalization is to learn a domain agnostic representation such that it will not be affected by some common covariate shifts on image (e.g., lightning, pose, background). Domain generalization-based methods can be further divided into four categories: domain adversarial learning (Ganin and Lempitsky (2015); Ganin et al. (2016); Li et al. (2018)), domain alignment (Tzeng et al. (2014); Wang et al. (2018); Pan et al. (2018)), kernel methods (Blanchard et al. (2011); Blanchard et al. (2021)) and feature normalization (CHEN et al. (2019); Sun and Saenko (2016)). Causal invariant representation learning (Arjovsky et al. (2020); Krueger et al. (2021); Pfister et al. (2018); Oberst et al. (2021); Heinze-Deml et al. (2018); Xie et al. (2021)) is another popular method in this domain. Compared to other methods, these addresses the OOD generalization problem in a more principled way by exploring causal variables and structures. Causal learning methods make some implicit assumptions: some underlying causal structure exists inside data and data is pooled from multiple training environments.

Lastly, optimization-based methods (Liang et al. (2020); Duchi et al. (2022); Hu et al. (2018); Kuhn et al. (2015); Xu and Yang (2013)) are also gaining popularity because of theoretical guarantees they can provide, and these methods are model and data structure agnostic. Distributionally Robust Optimization (DRO) approaches are extremely popular in this category. DRO methods directly solve the OOD generalization problem by optimizing the worst-case error over a training set. In the following Section 2.5, we will review GroupDRO (Liang et al. (2020)) method in detail.

### 2.4 Invariant Risk Minimization(IRM)

Invariant risk minimization (Arjovsky et al. (2020)) describes a practical and very elegant idea to solve a problem of out of distribution generalization, and it has become extremely

popular method in the OOD literature because it is one of the first methods that proposed to learn an invariant causal representation that can be applicable to deep learning models.

In IRM, authors used the concept of "Environments" and assumed that the training data is pooled from different environments, which have different data distributions and during test time this environment label is not available. The goal of IRM is to learn robust representations that are based on invariant causal association, rather than spurious correlations that might be present in the data across different environments. It is extremely difficult to extract these causal associations from i.i.d data only. Thus, data distributions differ in different environments but there is an underlying causal structure between label $\mathcal{Y}$ and some of feature's $\mathcal{X}$ that remains invariant across all environments.

The objective of IRM is to learn a model that is simultaneously Bayes optimal across all training environments. The above objective can be translated into a constrained optimization problem,

$$\begin{gathered} \min_{\substack{\Phi:\mathcal{X}\to\mathcal{H}, \\ w:\mathcal{H}\to\mathcal{X}}} \sum_{e\in E_{tr}} R^e(w\circ\Phi) \\ \text{subject to } w \in \text{argmin}_{\bar{w}:\mathcal{H}\to\mathcal{Y}}\, R^e(\bar{w}\circ\Phi),\ \text{for all } e\in E_{tr}. \end{gathered} \tag{2.2}$$

In this optimization problem, $R^e$ represents risk (loss) for environment $e$ and we are imposing a constraint on the data representation $\phi$, such that the classifier $w$, which takes this data representation as an input, will be simultaneously Bayes optimal across the set of all training environments ($E_{tr}$). In that paper, authors present a hypothesis that it would only be true if the data representation is built from an invariant and causal variable which does not change across different environments or under distributional shifts. The authors have provided detailed proof for this hypothesis in their work (Arjovsky et al. (2020)).

However, this bi-level optimization problem is computationally intractable, Thus the authors propose a relaxation of the problem by introducing a new invariance penalty for a single level optimization,

$$\min_{\Phi:\mathcal{X}\to\mathcal{Y}} \sum_{e\in E_{tr}} \left(R^e(\Phi) + \lambda \cdot \|\nabla_w|_{w=1.0} R^e(w\cdot\Phi)\|_2^2\right). \tag{2.3}$$

There are many interesting follow up papers (Lu et al. (2022); Ahuja et al. (2020); Ahuja et al. (2021); Rosenfeld et al. (2021); Kamath et al. (2021)) based on the IRM. Recently, Rosenfeld et al. (2021) showed some limitations of the IRM approach and proved that this method will fail when the number of training environments is exceedingly small.

### 2.5 Group Distributionally Robust Optimization (GroupDRO)

As noted in the introduction, overparameterized deep learning models are susceptible to absorbing spurious correlations during training with empirical risk minimization (ERM). To address this issue, various optimization-based methods have been proposed to reduce the reliance on spurious correlations. One such popular method in this research direction is Distributionally Robust Optimization (DRO) applied at the group level. By considering the distribution of examples across multiple groups rather than individually, GroupDRO (Liang et al. (2020)) aims to improve the model's robustness to distributional shifts and mitigate the effects of spurious correlations.

Group distributional robust optimization (GroupDRO) is a method that leverages prior information about spurious factors to divide the data into groups. For instance, in the context of a camel-cow classification task where all the images are captured in either a desert or a meadow, the background label can only take on two values. Given that this is a binary classification task, the training data can be partitioned into a total of four groups, representing the combination of background and class. During training, the loss

can be separately evaluated for each group, and the primary objective of GroupDRO is to minimize the empirical worst-case group loss,

$$\begin{aligned} \theta^* &= \arg\min_{\theta \in \Theta} R(\theta), \\ R(\theta) &:= \sup_{Q \in \mathcal{Q}} \mathbb{E}_{(x,y)\sim Q} \left[ \ell(\theta; (x, y)) \right]. \end{aligned} \tag{2.4}$$

In the remainder of the paper, the authors demonstrate how optimizing the worst-case group loss can aid in improving the generalization of overparameterized models against spurious correlations. Additionally, the authors present several generalization bounds and convergence properties of this algorithm. The authors further demonstrate how group-wise optimization can enhance accuracy in minority groups, mitigate bias, and promote fairness.

### 2.6 Variance Risk Extrapolation (VREx)

In recent years, VReX has become increasingly popular in the literature on out-of-distribution (OOD) learning. VReX, as presented in Krueger et al. (2021), can be regarded as a generalized version of IRM (Arjovsky et al. (2020)) and has connections with the GroupDRO (Liang et al. (2020)) method. In this work, the authors utilize environmental assumptions, whereby the training data is pooled from various training distributions/environments.

In VReX, the authors propose two penalty formulations to promote distributional robustness. The first version, known as MM-REx (Minimax Risk Extrapolation), can be regarded as an extension of the Group DRO method. This method can extrapolate or provide theoretical guarantees for robustness outside the convex hull of training distributions, as illustrated in Figure 2.1. However, in practice, this method has been shown to be very unstable. The primary objective of MM-REx is to optimize over a perturbation set comprising various affine combinations of training distributions,

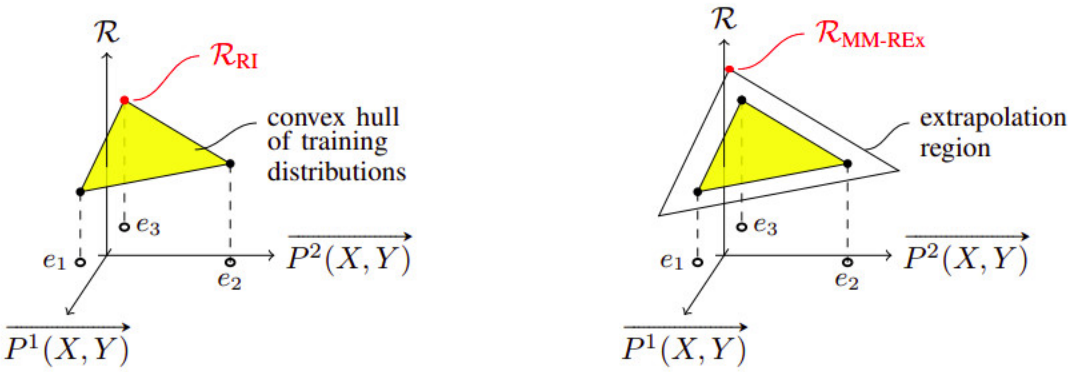



Figure 2.1: Left: Robust optimization optimizes worst-case performance over the convex hull of training distributions. Right: By extrapolating risks, REx encourages robustness to larger shifts (adapted from Krueger et al. (2021)).

$$\text{RMM-REx}(\theta) = \max_{\sum_e \lambda_e = 1, \lambda_e \geq \lambda_{\min}} \sum_{e=1}^{m} \lambda_e \text{Re}(\theta) = (1 - m\lambda_{\min}) \max_e \text{Re}(\theta) + \lambda_{\min} \sum_{e=1}^{m} \text{Re}(\theta). \tag{2.5}$$

Here $m$ is the number of domains, and the hyperparameter $\lambda_{\min}$ controls the extrapolation. For negative values of $\lambda_{\min}$, MM-REx (Minimax Risk Extrapolation) puts negative weights on the risk of all but the worst-case domain, and as $\lambda_{\min} \to -\infty$, this criterion enforces strict equality between training risks. Nevertheless, negative coefficients allow extrapolating to the more extreme variations or distribution shifts of the convex hull of training distributions. A hyper parameter $\lambda_{\min}$ also has a geometrical significance: larger values of $\lambda_{\min}$ expand the perturbation set away from the convex hull of risk/loss of different training distributions and promote a flatter risk-plane, which helps in better generalization.

In contrast, VREx (Variance Risk Extrapolation) is more stable, simpler, and practical;

$$\text{RV-REx}(\theta) = \beta \, \text{Var}(R_1(\theta), \ldots, R_m(\theta)) + \sum_{e=1}^{m} R_e(\theta). \tag{2.6}$$

Here $\beta \in [0, \infty)$ controls the tradeoff between reducing average risk and enforcing equality of risk: $\beta = 0$ recovers ERM, while $\beta \to \infty$ leads VREx to focus entirely on making the risks equal (minimizing the variance of risks).

Later in the paper, the authors establish a connection between their proposed method and causal inference, and provide a comparative analysis with the IRM and GroupDRO approaches. The results demonstrate the superior performance of the proposed method over both IRM and GroupDRO in terms of providing covariate shift robustness and invariant prediction.

### 2.7 Contrastive Learning

Recently, Contrastive learning (Zbontar et al. (2021); Caron et al. (2021); Song et al. (2015); Gutmann and Hyvärinen (2010); Chen et al. (2020); Khosla et al. (2021); Fan et al. (2020)) has emerged as one of the most powerful methods in self-supervised learning. Some of the recent advances in self-supervised visual representation learning such as SimCLR (Chen et al. (2020)), CLIP (Radford et al. (2021)), DALLE-2 (Ramesh et al. (2022)) have been powered by contrastive learning techniques. The objective of contrastive learning is to pull the representation of an anchor/target point closer to the representation of certain points, known as 'positives', while simultaneously pushing it away from other points, known as 'negatives' in an embedding space. The above objective can be translated into the following loss function as described in Chen et al. (2020),

$$L_{SimCLR} = -\frac{1}{N} \sum_{i=1}^{2N} \sum_{j=1}^{2N} \log \frac{\exp(\text{sim}(z_i, z_j)/\tau)}{\sum_{k=1}^{2N} 1_{[k \neq i]} \exp(\text{sim}(z_i, z_k)/\tau)}. \tag{2.7}$$

In self-supervised contrastive learning (Chen et al. (2020)), the augmentation module applies two different augmentation operations to each image in a minibatch of N samples, resulting in a total of 2N samples being generated. In the above loss function, $N$ is the

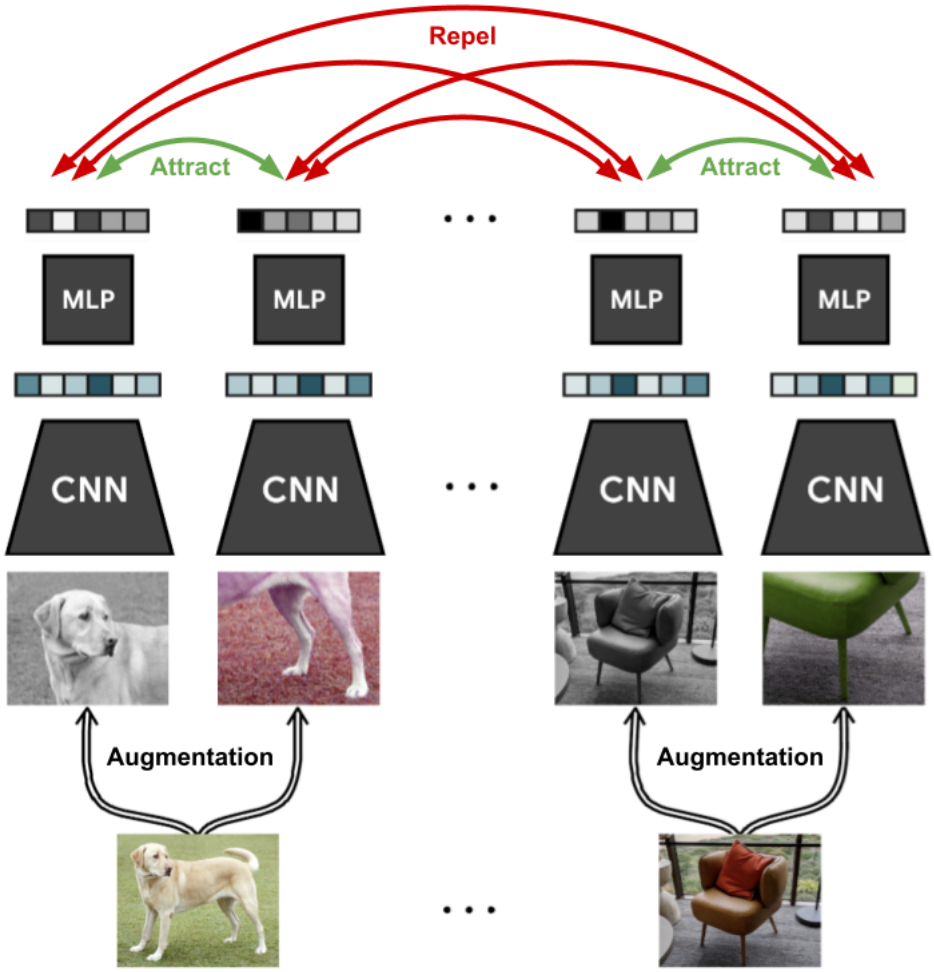



Figure 2.2: An illustration of SimCLR framework (adapted from Chen et al. (2020)).

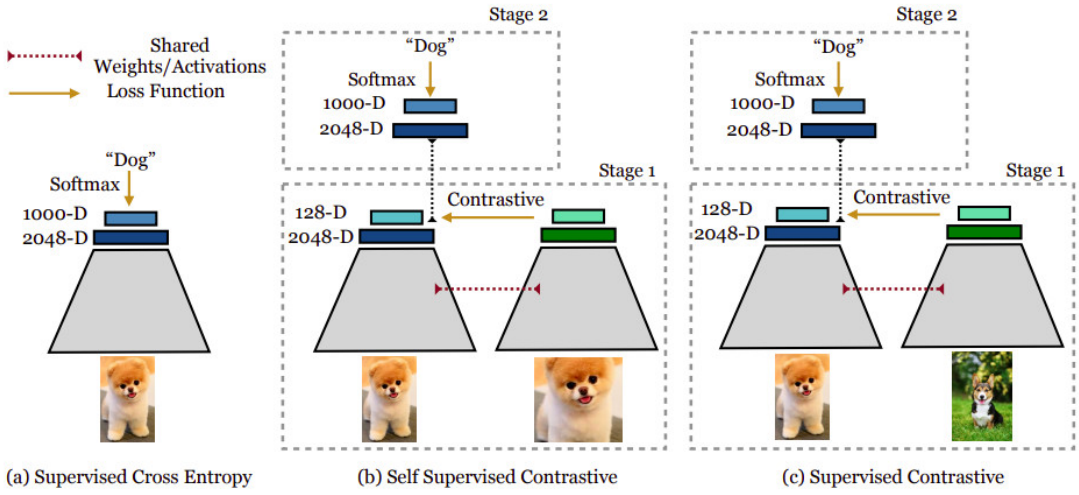



Figure 2.3: (a) Training a image classifier using cross entropy loss (b) Self supervised contrastive learning uses augmentation to learn a representation (c) Supervised contrastive learning uses contrastive learning and label information to improve the learned representation (adapted from Khosla et al. (2021)).

batch size, $z_i$ and $z_j$ are the representations of the same image under two different augmentations, $\text{sim}(z_i, z_j) = \frac{z_i^T z_j}{|z_i||z_j|}$ is the cosine similarity between the two representations, $\tau$ is a temperature parameter that controls the concentration of the distribution, and $1_{[k \neq i]}$ is an indicator function that equals 1 if $k \neq i$ and 0 otherwise.

It turns out that contrastive learning is a good training strategy for deep learning models. Chen et al. (2020) showed that a self-supervised contrastive learning method can learn a representation that matches the performance of supervised methods. Recently, Khosla et al. (2021) indicated that utilizing supervised contrastive learning, where positive-negative pairs are formed based on class labels, could potentially be more effective for training models in general.

In Figure 2.3, a visual comparison is shown between different learning methods. Image (a) represents a normal training of a deep learning model using cross entropy loss. Image (b) represents self-supervised contrastive learning framework, and (c) represents supervised contrastive learning framework. Note that both (b) and (c) are divided into two stages of

training. In the first stage, only the encoder is trained with contrastive loss and in the second stage, the classifier model is trained with cross entropy loss function and training for the encoder is frozen.

### 2.8 GradCAM

GradCAM (Gradient-weighted Class Activation Mapping)(Selvaraju et al. (2019)) is a popular explainable AI method to visualize and understand the decision process of a deep learning model. It is a class specific activation method, and it provides insights into which regions of an input image are important for a CNN's decision-making process for predicting a particular class, by generating a heatmap over the region.

GradCAM computes the gradients of the target class score with respect to the feature maps in the last convolutional layer of a neural network. It then aggregates the gradients and produces a weighted map, which is used to generate the final heatmap. This approach is efficient and computationally inexpensive, making it ideal for use in a wide range of computer vision tasks.

GradCAM is both model and task agnostic, thus it can be used for variety of computer vision tasks such as object detection, classification, and image segmentation, and it can be

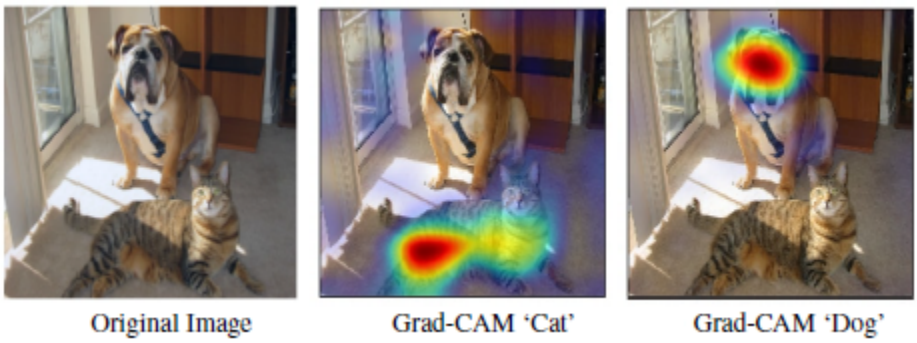



Figure 2.4: Ground truth image(left), Class activation heatmap generated by GradCAM to highlight discriminative regions for class 'cat' (middle), and Class activation heatmap to highlight discriminative regions for class 'dog' (right) (adapted from Selvaraju et al. (2019)).

used for any deep learning architecture. It has been shown to improve model interpretability which is extremely important for some important computer vision applications such as medical imaging and self-driving technology.

# Chapter 3

# Experimental setup

This chapter introduces the experimental setting for training and evaluating existing OOD methods, namely IRM (Arjovsky et al. (2020)), GroupDRO (Liang et al. (2020)), and VREx (Krueger et al. (2021)). It is challenging to compare these methods in a fair setting due to differences in datasets, models, and hyperparameters as noted in their respective papers. The chapter is organized as follows: Section 3.1 introduces datasets for benchmarking these methods, while Section 3.2 discusses the experimental setup. Section 3.3 presents some preliminary experimental findings on the CMNIST dataset, and Section 3.4 focuses on the findings of experiments conducted on the Waterbirds and CelebA datasets. Lastly, Section 3.5 discusses the visual comparison of these OOD methods using explainable AI method.

## 3.1 Datasets

1. **CMNIST:**

   CMNIST is a synthetic dataset introduced by Arjovsky et al. (2020). The dataset is derived from MNIST, which is a well-known and widely used dataset of handwritten digits commonly used for image recognition tasks. It consists of a set of 70,000 grayscale images, each of which is 28x28 pixels in size. These images depict handwritten digits ranging from 0 to 9. The dataset is split into two parts: a training set of 60,000 images and a test set of 10,000 images.

   CMNIST is a well known dataset in OOD lieturatue and is commonly used for a binary classification task, as describe in Arjovsky et al. (2020). The first 5 digits (0-4) are labeled as 0 and the remaining digits are labeled as 1. Additionally, images are

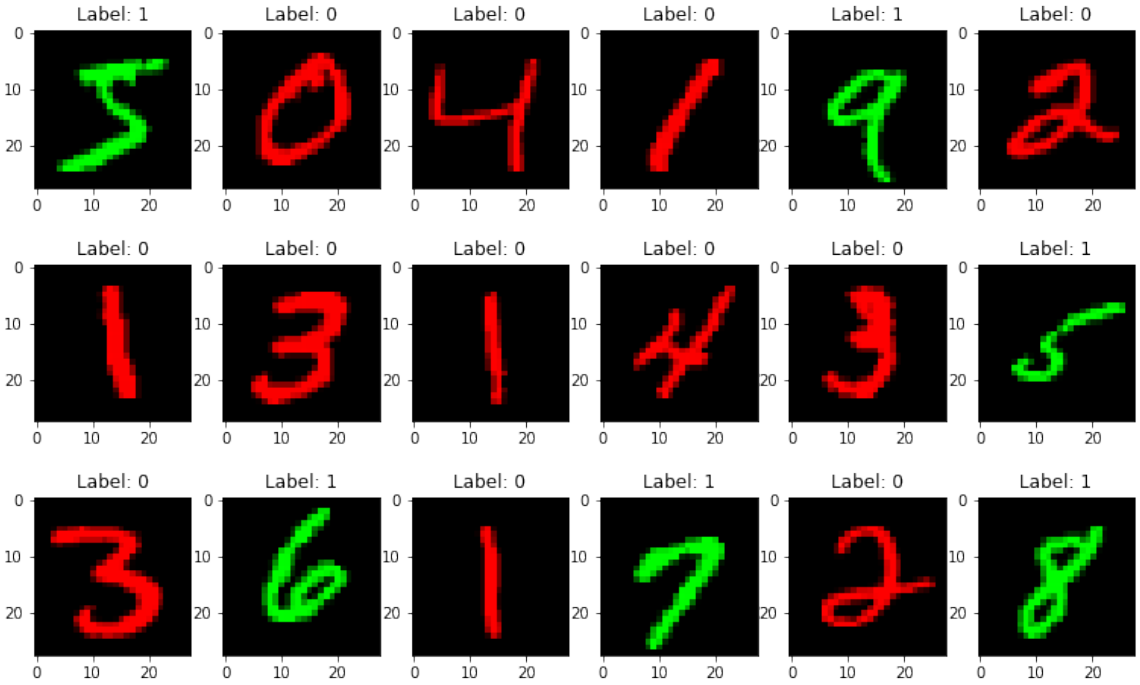



Figure 3.1: Sample images from the CMNIST Dataset (adapted from Arjovsky et al. (2020)).

colored either red or green in such a way that it spuriously correlated with the target label. Thus, during training there is a strong correlation between color and target label. However, during test time, to generate a distribution shift, the color is flipped according to some probability distribution.

2. **Waterbirds:**

   The Waterbirds dataset was introduced by Liang et al. (2020), and is a binary bird classification task. The dataset is constructed by combining two datasets: Caltech-UCSD Birds (CUB) 200-2011 (Wah et al. (2011)) and Places (Zhou et al. (2017)). Caltech Birds is a popular benchmarking dataset for bird species classification. CUB contains around 11,788 images of 200 bird species, and each image is associated with various attributes (order, family, species etc.) and pixel level segmentation mask. Places (Zhou et al. (2017)) is a benchmarking dataset containing 10 million scene

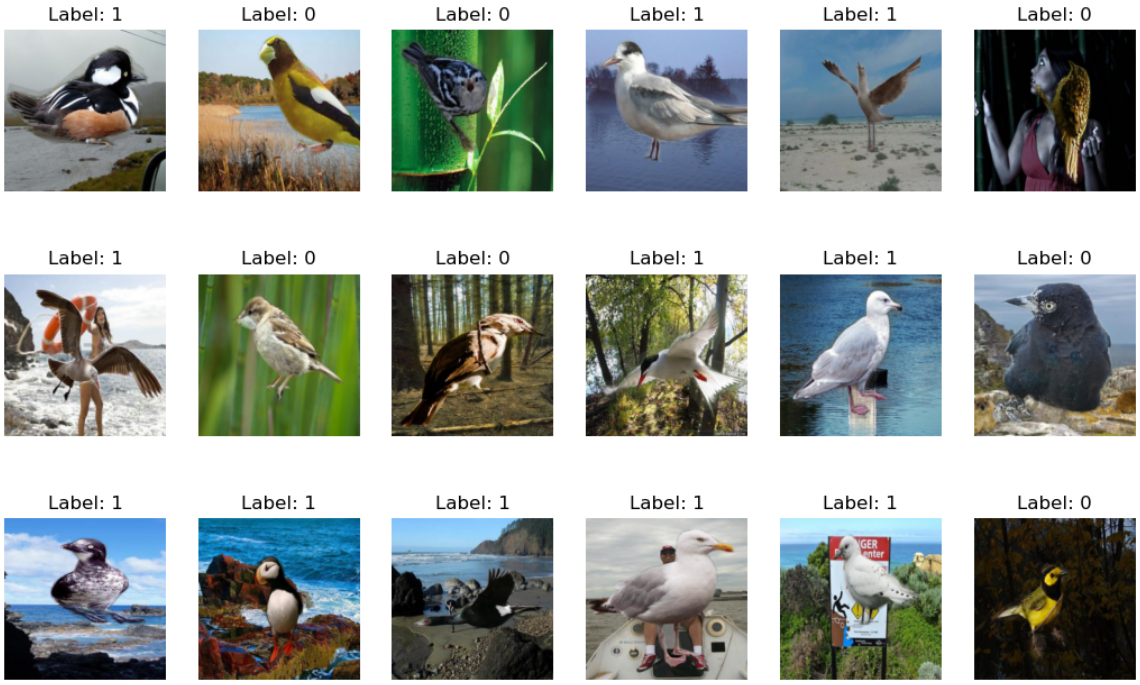



Figure 3.2: Sample images from the Waterbirds dataset (adapted from Liang et al. (2020)).

photographs. Images are labeled with different types of environments encountered in the world, such as Nature, Urban, Indoor, etc.

To generate a Waterbirds dataset, authors first queried both CUB (Wah et al. (2011)) and Places (Zhou et al. (2017)) datasets with specific attributes to create two sets $\mathcal{Y}$ and $\mathcal{A}$. $\mathcal{Y}$ is a set of birdtypes (waterbirds and landbirds) and $\mathcal{A}$ is a set of landmarks (water background or land background). First, CUB (Wah et al. (2011)) dataset was queried with specific species, and after those species labels are mapped to binary label to create a set $\mathcal{Y}$ - all the seabird (albatross, auklet, cormorant, frigatebird, fulmar, gull, jaeger, kittiwake, pelican, puffin, or tern) and waterfowl (gadwall, grebe, mallard, merganser, guillemot, or Pacific loon) species are mapped to 'waterbird' label and remaining species are mapped to the 'landbird' label. To create a set $\mathcal{A}$, the Places (Zhou et al. (2017)) dataset is queried by two types of scene background (water background and land background). After that, based on the segmentation

mask information, each image from set $\mathcal{Y}$ is cropped from its original background and pasted onto an image from set $\mathcal{A}$. Now the images can be divided into four groups based on the bird type and the background. The original training and test set are resampled to create a new balanced training and test set. The training set contains 1600 images (864 waterbirds and 736 images of landbirds), and the test set contains 2000 images (1000 images of each class). Now in this dataset, the target label is spuriously correlated with landmark : in the training dataset, around 80% of waterbirds are against water backgrounds, and about 90% of landbirds are against a land background. However, during test time, this correlation does not exist, and each class has an equal number of images for both backgrounds.

3. **CelebA:**

   CelebA (Liu et al. (2015)) is a large-scale face attributes dataset with more than 200K celebrity images, and each image has 40 facial attribute annotations. Images have a significant diversity and cover large variations in background and pose. This dataset is extremely popular for a large set of computer vision tasks such as face detection, recognition, and localization.

   To use this dataset for out of distribution generalization, the dataset is first queried by the hair color attribute ($label = \{Dark, Blond\}$) for the target variable and gender attributes ($gender = \{male, female\}$) to create a smaller subset of the original dataset. The goal of using this dataset is to understand the effect of demographic information such as gender and ethnicity, which is quite common in real world applications of computer vision. The objective is to predict the hair color of a celebrity from the image. The original dataset contains 1627770 images. However due to the limitations of compute and resources, the dataset is filtered to create new train and test datasets. The filtered training dataset contains 7200 images and the test set contains 5548 im-

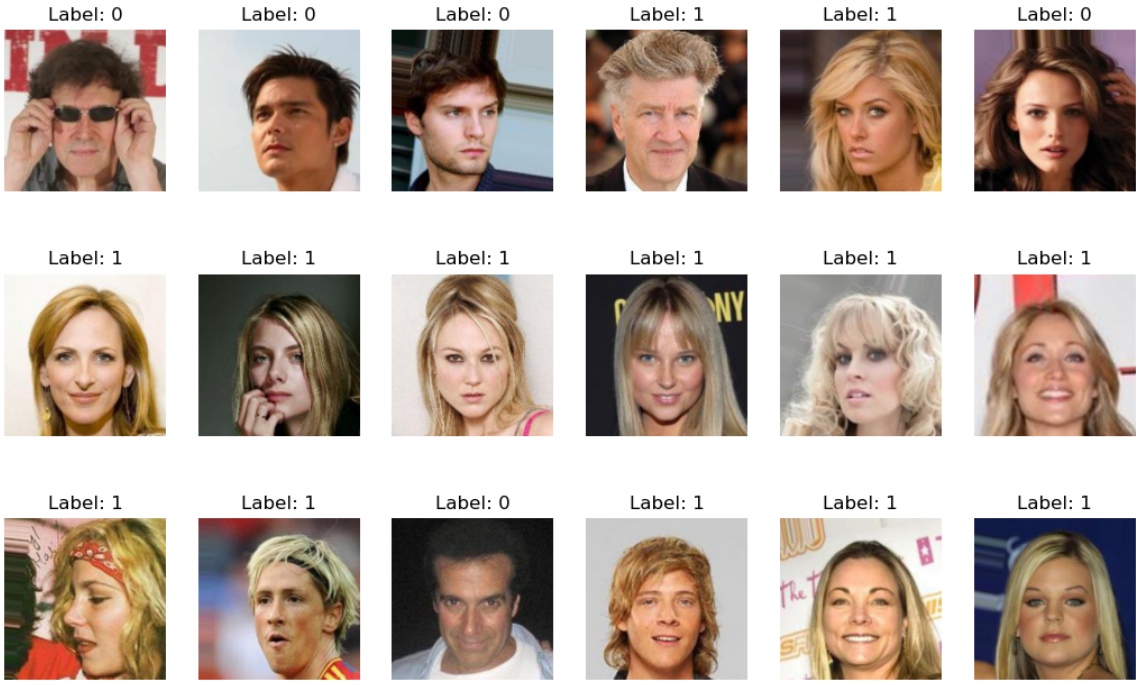



Figure 3.3: Sample images from the CelebA dataset (adapted from Liu et al. (2015)).

ages. In the training dataset, the gender attribute is spuriously correlated with the hair color: 87% of men have blond hair and 87% of women have dark hair. However, this correlation does not exist in the test set, as each class has an equal proportion of images for both genders.

### 3.2 Experimental setting

As discussed in the introduction, the evaluation of out-of-distribution (OOD) methods is widely acknowledged as a challenging issue in the literature. Establishing a benchmark is therefore crucial to enable a fair comparison of these methods. In this study, we utilized the CMNIST dataset introduced by Arjovsky et al. (2020) as a starting point. To facilitate the training of IRM and VREx methods, the training dataset containing 40,000 images was divided into two disjoint sets to create two separate environments. During the creation of this dataset, labels were initially flipped with 25% probability to create a base dataset.

This base dataset was later modified to create a set of environments. To create the first training environment, colors were randomly flipped with 20% probability, and to create the second training environment, colors were randomly flipped with 10% probability to simulate a distribution shift. In the test set, colors were flipped with 90% probability. For training the GroupDRO method, data from these environments were combined, and group labels were inferred from the prior information about labels and colors during training.

As a model architecture, a simple 2-layer CNN is used as a starting point. All the other hyperparameters are kept constant during training and testing for all three methods, with ERM serving as a baseline. The models were trained for 100 epochs using the Adam optimizer, binary cross-entropy loss, a learning rate of 0.001, and a batch size of 64. To train an IRM model, the most important technique, as mentioned in the Arjovsky et al. (2020), was to dynamically increase the gradient norm penalty as a function of epochs. Furthermore, for the VREx model, as stated in the original paper (Krueger et al. (2021)) and implementation, a remarkably high penalty weight of 10000 was employed. The results of this experiment are presented in the subsequent section.

In a separate experiment, the ResNet50 model was implemented in lieu of the vanilla CNN model. All other hyperparameters were kept consistent with the previous experiment. The outcomes of this experiment are elaborated upon in the subsequent section.

### 3.3 Preliminary Results

The evaluation of all the aforementioned methods was performed based on the accuracy obtained on the OOD test set. The results of the first experiment are presented in Table 3.1.

The obtained results demonstrate that the model trained with ERM exhibits inadequate performance on the OOD test set, with the ERM baseline accuracy being a mere 10%.

| Method name | Train accuracy | OOD test accuracy |
|---|---|---|
| ERM(baseline) | 85.2 | 10.34 |
| IRM | 71.09 | 65.17 |
| GroupDRO | 71.41 | 72.1 |
| VREx | 65.3 | 72.60 |

Table 3.1: Accuracy analysis of Experiment 1: Creating a benchmark to compare different OOD methods on CMNIST dataset with 2 layer vanilla CNN model.

Conversely, the other methods produce substantially better outcomes than ERM, with IRM exhibiting approximately 55% accuracy gain in comparison to ERM. GroupDRO and VREx exhibit around 62% accuracy gain relative to the baseline. Additionally, it is noteworthy that the training accuracy for OOD methods is significantly lower than that of ERM, which aligns with prior research findings (Tsipras et al. (2019)).

| Method name | Train accuracy | OOD test accuracy |
|---|---|---|
| ERM(baseline) | 86.33 | 16.4 |
| IRM | 79.5 | 45.39 |
| GroupDRO | 84.2 | 56.1 |
| VREx | 87.04 | 44.2 |

Table 3.2: Accuracy analysis of Experiment 2: Creating a benchmark to compare different OOD methods on CMNIST dataset with ResNet50 model.

Table 3.2 presents the results of the second experiment. The results show that training a larger model improves training accuracy for all methods, and the test accuracy for the baseline also increases by 6%. However, the OOD test accuracy significantly decreases for the OOD methods. This may be attributed to the fact that ResNet50 is an overparameterized model for this toy dataset, which causes the model to memorize spurious correlations, leading to a drop in accuracy.

### 3.4 Creating a benchmark

In this section, we present the methodology for creating an evaluation benchmark to compare the performance of different methods on two datasets: Waterbirds and CelebA. The

experimental settings were kept the same for both datasets, with the aim of conducting a fair comparison.

To train the models, we used the ResNet50 model architecture and trained the model using binary cross-entropy loss for 100 epochs with the Adam optimizer. The initial learning rate was set to 1e-4, the batch size was set to 128, and the weight decay was set to 1e-3.

For the Waterbirds dataset, we created two training environments, each containing a disjoint set of 800 images from the training set of 1600 images. In the first training environment, the class labels and background were strongly correlated, with 90% of the Waterbirds images in water background and the landbirds in the land background. In the second training environment, this correlation was reduced to 0.85. We trained the IRM and VREx models on these two environments.

Similarly, for the CelebA dataset, we divided the training dataset, which contained a total of 7200 images, into two environments for training the IRM and VREx models. The spurious correlation between the target variable and the nuisance (gender in this case) was kept the same as in the Waterbirds dataset. To train the GroupDRO model, the group label was generated from the nuisance (background label in Waterbirds and gender label in CelebA) and target variable.

To create an OOD test set for both datasets, we ensured that there was no correlation between the target label and nuisance, as the joint distribution of them was uniform.

In the following section, we present the results for both the Waterbirds and CelebA datasets, with the aim of providing insights into the performance of the different methods on these datasets. Our methodology provides a standardized approach to evaluate the effectiveness of these methods, which can be used as a benchmark for future research in the following sections.

### 3.5 Results and findings

This section presents the results of the latest experiment conducted on both datasets. The evaluation metric used for assessing the model's performance on the out-of-distribution (OOD) test set is accuracy. Furthermore, to gain a more comprehensive understanding of the algorithms, we will also calculate group-wise accuracy. The group-wise accuracy analysis will enable a detailed examination of the algorithms' performance on different subgroups within the dataset.

| **Method name** | **Train accuracy** | **OOD test accuracy** |
|---|---|---|
| ERM(baseline) | 84.81 | 69.75 |
| IRM | 91.0 | 75.5 |
| GroupDRO | 98.9 | 79.5 |
| VREx | 94.29 | 80.6 |

Table 3.3: Evaluation benchmark for OOD methods on the Waterbirds dataset.

The evaluation of various methods on the Waterbirds dataset is presented in Table 3.3. The accuracy on the test set demonstrates that the methods outperformed the ERM (baseline) model by a margin of around 6-10%. To gain a more detailed understanding of the models' performance, we calculated group-wise accuracy. The results from Table 3.4 indicate that the ERM model performs better on majority groups, i.e., a set of examples where the target label is spuriously correlated with the background. However, its accuracy significantly degrades for minority groups, i.e., a set of examples over which the spurious correlation does not hold. In contrast, other methods show better performance than the baseline on the minority groups.

The results on the CelebA dataset are presented in Tables 3.5 and 3.6. The IRM method performed poorly compared to the ERM method on both the train and test sets. In contrast, the VRex and GroupDRO methods outperformed the ERM method with an accuracy gain of around 2%. However, calculating the group-wise accuracies revealed a different story. Table

| Method name | Majority groups | | Minority groups | |
|---|---|---|---|---|
| | Group 1 | Group 2 | Group 3 | Group 4 |
| ERM(baseline) | 95.4 | 83.2 | 38 | 62.4 |
| IRM | 94.4 | 90.0 | 49.4 | 68.2 |
| GroupDRO | 92.6 | 87.8 | 62.6 | 76.0 |
| VREx | 92.0 | 92.0 | 71.8 | 66.8 |

Table 3.4: Groupwise accuracy analysis of different OOD methods on the Waterbirds dataset. Group 1 (waterbirds on water) and Group 2 (landbirds on land) are considered as majority groups( background and target label strongly correlated during training), Group 3 (landbirds on water) and Group 4 (waterbirds on land) are considered as minority groups.

3.6 presents the group-wise accuracy for the different methods, indicating that GroupDRO is the only method to outperform the ERM baseline on all minority groups with a significant gain.

| Method name | Train accuracy | OOD test accuracy |
|---|---|---|
| ERM(baseline) | 91.2 | 87.83 |
| IRM | 89.87 | 87.54 |
| GroupDRO | 94.99 | 89.63 |
| VREx | 93 | 89.54 |

Table 3.5: Evaluation benchmark for OOD methods on the CelebA dataset.

| Method name | Majority groups | | Minority groups | |
|---|---|---|---|---|
| | Group 1 | Group 2 | Group 3 | Group 4 |
| ERM(baseline) | 92.29 | 98.05 | 75.63 | 85.36 |
| IRM | 95.17 | 98.13 | 77.87 | 79.02 |
| GroupDRO | 91.35 | 93.44 | 85.58 | 88.18 |
| VREx | 94.09 | 96.68 | 81.33 | 82.26 |

Table 3.6: Groupwise accuracy analysis of different OOD methods on the CelebA dataset. Group 1 (Dark-haired men) and Group 2 (Blond-haired women) are considered as majority groups, Group 3 (Dark-haired women) and Group 4 (Blond-haired men) are considered as minority groups.

### 3.6 Visualization

Calculating the average and group-wise accuracy is a commonly used metric for comparing the performance of various out-of-distribution (OOD) methods. However, this metric fails

to provide insights into whether the model is learning robust, invariant, and task-specific features. In order to test the efficacy of the model's learned features, we propose the use of an explainable artificial intelligence (XAI) method called Gradient-weighted Class Activation Mapping (GradCAM). By using GradCAM, we can visualize the regions where the model is focusing to make a prediction. Specifically, we are interested in examining the visual attention of the model towards minority groups, where spurious correlations between the target label and background are less likely to hold. If the model is found to be focusing more on non-semantic background features, it indicates that it may not learning generalizable features, rendering it susceptible to failure under distribution shifts.

Table 3.7 illustrates the class activation maps generated by applying GradCAM to models trained using various OOD methods on Waterbirds dataset. The results indicate that models trained with ERM and IRM tend to focus slightly on non-robust background features that are spuriously correlated with the target label during training. Conversely, models trained with VREx and GroupDRO methods demonstrate a propensity to learn task-specific features and emphasize on the semantic features of birds while making predictions about the class.

Table 3.8 presents the class activation maps obtained by utilizing the GradCAM technique on models trained using distinct OOD methods on the CelebA dataset. During training on this dataset, there exists a spurious correlation between face color and the target label, i.e., hair color. The outcomes indicate that models trained with ERM and IRM approaches tend to focus on skin color and facial characteristics that are spurious to the target label during training. In contrast, models trained with VREx and GroupDRO methods manifest a tendency to acquire task-specific features and prioritize hair color while making predictions about the class.

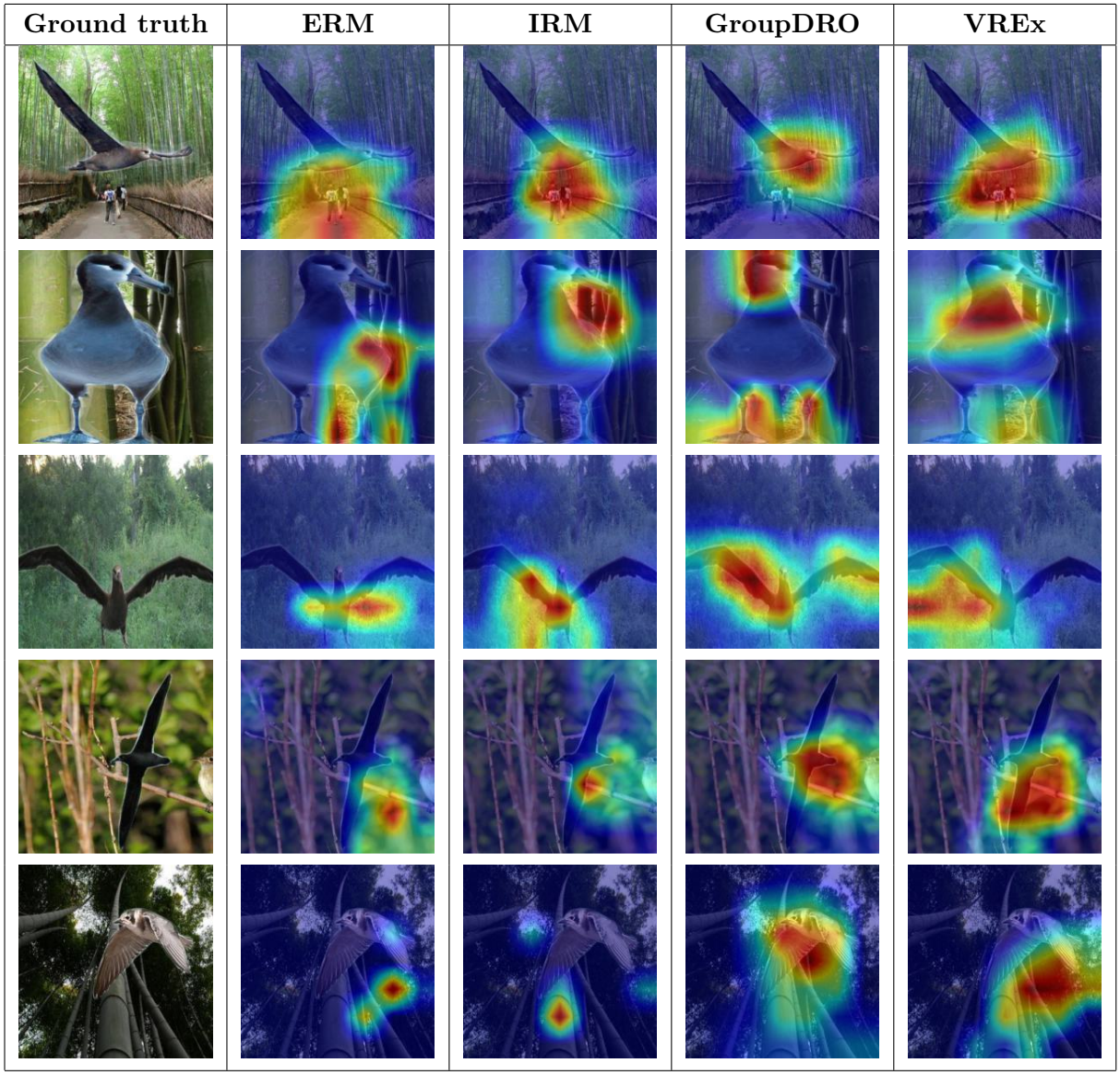



Table 3.7: Class activation heatmaps generated on the Waterbirds test set to highlight discriminative features learned by different OOD methods.

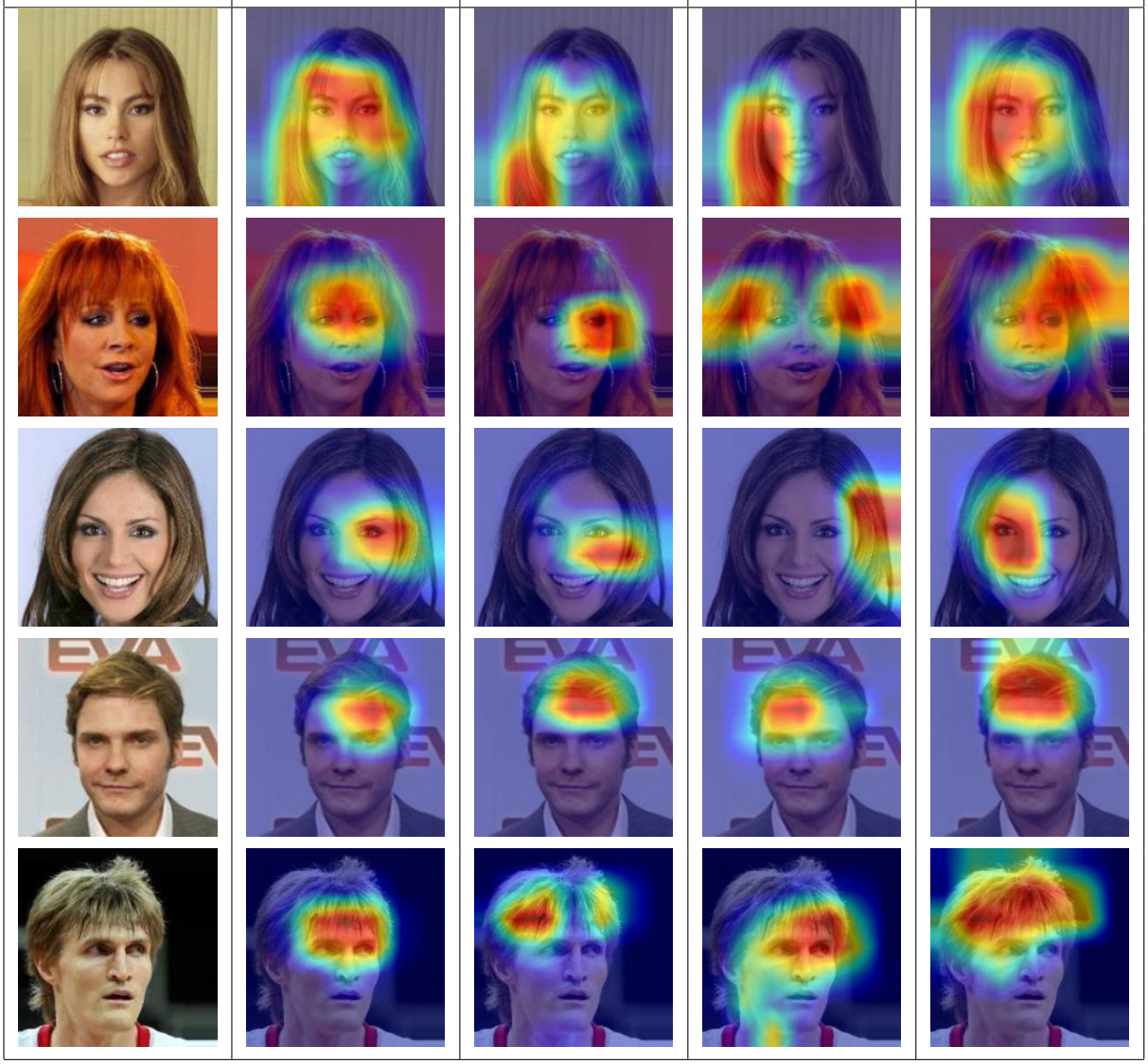
| Ground truth | ERM | IRM | GroupDRO | VREx |
|---|---|---|---|---|
|  | | | | |

Table 3.8: Class activation heatmaps generated on the CelebA test set to highlight discriminative features learned by different OOD methods.

## 3.7 Conclusion

Based on the experimental results, it is evident that the Group Distributionally Robust Optimization (GroupDRO) method is a highly effective technique in comparison to other OOD generalization methods used in the benchmark. In the subsequent section, we delve deeper into the efficacy of the GroupDRO method and investigate the key ingredient that drives its success, namely, group-based prior information. Further experiments are conducted to gain a comprehensive understanding of the GroupDRO method's effectiveness in dealing with OOD scenarios.

# Chapter 4

# Group wise optimization and reweighted sampling

## 4.1 Introduction

In this section, we will discuss the importance of group-based optimization and group wise reweighted sampling. Weighted sampling is a very popular method in machine learning literature to sample data points from a dataset with non-uniform probability mass assigned to each point. It is often used when the dataset is highly imbalanced or skewed. Group wise reweighted sampling is a slightly modified version of weighted sampling. Here the weights represent the distribution of groups in the training dataset. During training, group labels are available, and based on the information of group labels, the weights can be calculated for group wise reweighted sampling. In our previous experiment, we observed that the GroupDRO method outperformed other methods on both the waterbirds and CelebA datasets. However, upon analyzing the original code provided by Liang et al. (2020), we discovered that the authors had used group-based reweighted sampling before conducting the group-based optimization, which was the main objective of their method. To explore the impact of this technique, we conducted an experiment where we removed the group-based reweighted sampling and repeated the experiment on both the Waterbirds and the CelebA datasets. The results of this experiment were quite astonishing and will be discussed in the table below.

| **Method name** | **Waterbird test set** | **CelebA test set** |
|---|---|---|
| GroupDRO | 79.5 | 89.63 |
| GroupDRO without sampling | 66.66 | 86.38 |

Table 4.1: Comparing GroupDRO with and without groupwise reweighted sampling

Based on the findings presented in Table 4.1, it can be observed that GroupDRO exhibits a significant decrease in performance when it is trained without groupwise reweighted sampling. On the Waterbirds dataset, there is a notable accuracy drop of approximately 12%. Conversely, on the CelebA dataset, the decline in accuracy is relatively insignificant, at around 3%. These results suggest that groupwise reweighted sampling is a crucial step in groupwise optimization. In the subsequent section, we will conduct an experiment to investigate whether implementing groupwise reweighted sampling can enhance the performance of the empirical risk minimization (ERM) for out-of-distribution (OOD) generalization task.

### 4.2 Group wise reweighted sampling for improving ERM

The aim of this experiment was to examine the impact of group-based reweighted sampling on the generalization capabilities of empirical risk minimization (ERM) on minority samples. A groupwise reweighted sampling strategy was applied to the baseline model with the ERM objective. All other hyperparameters were maintained consistent with the benchmarking setting.

| **Method name** | **Waterbird test set** | **CelebA test set** |
|---|---|---|
| ERM | 69.75 | 87.83 |
| ERM with sampling | 80.55 | 89.58 |
| IRM | 75.5 | 87.54 |
| GroupDRO | 79.5 | 89.63 |
| VREx | 80.6 | 88.59 |

Table 4.2: Comparision of ERM trained with groupwise reweighted sampling versus other OOD methods.

Table 4.2 presents the results that indicate the effectiveness of the groupwise reweighting strategy in improving the generalization of the model on the out-of-distribution test set. The experimental results on the Waterbirds dataset demonstrate that the model trained with groupwise reweighted sampling outperforms the vanilla ERM baseline by 11% in terms

of accuracy. Similarly, on the CelebA dataset, the model trained with groupwise reweighted sampling achieves a 2% accuracy gain compared to the ERM baseline. These findings underscore the significance of training a model with equal weightage on all types of groups to learn generalizable features that can better handle distribution shifts.

In the subsequent experiments, we plan to utilize this sampling strategy to train models with different out-of-distribution methods and investigate their performance on the test set.

### 4.3 Group wise reweighted sampling for improving existing OOD methods

In this study, we aim to investigate the effectiveness of group-wise reweighted sampling methods in improving the generalization capability of two existing out-of-distribution (OOD) methods: Invariant Risk Minimization (IRM) and Variance Risk Extrapolation (VREx). Since IRM and VREx rely on environmental assumptions of the training dataset, group-based weights will be estimated separately for each environment prior to applying sampling. All other parameters will be held constant to maintain consistency with the benchmark.

| **Method name** | **Waterbird test set** | **CelebA test set** |
|---|---|---|
| IRM | 75.5 | 87.54 |
| IRM with sampling | 81.15 | 89.70 |
| VREx | 80.6 | 88.59 |
| VREx with sampling | 88.59 | 89.41 |

Table 4.3: Results for applying groupwise reweighted sampling for other OOD methods.

Table 4.3 displays the results of our study, indicating an approximate 5% accuracy gain for IRM, 8% accuracy gain for VREx on Waterbirds dataset, and around a 1-2% accuracy gain on CelebA dataset following training with group-wise reweighted sampling. The observed increase in performance underscores the effectiveness of this approach for enhancing out-of-distribution generalization, affirming that prior knowledge of groups is a powerful assumption in the OOD literature. In the next section, we will examine a novel

method, 'Supervised Groupwise Contrastive Learning,' designed to improve representation for OOD generalization.

### 4.4 Visualization

Table 4.4 displays the visual representation of class activation mappings for comparing two models: vanilla ERM and ERM trained with a group-wise reweighted sampling strategy. The results indicate that the model trained with group-wise reweighted sampling acquires a more robust representation of the data and emphasizes on the task-specific features instead of the background. Therefore, this model exhibits superior performance against distribution shifts.

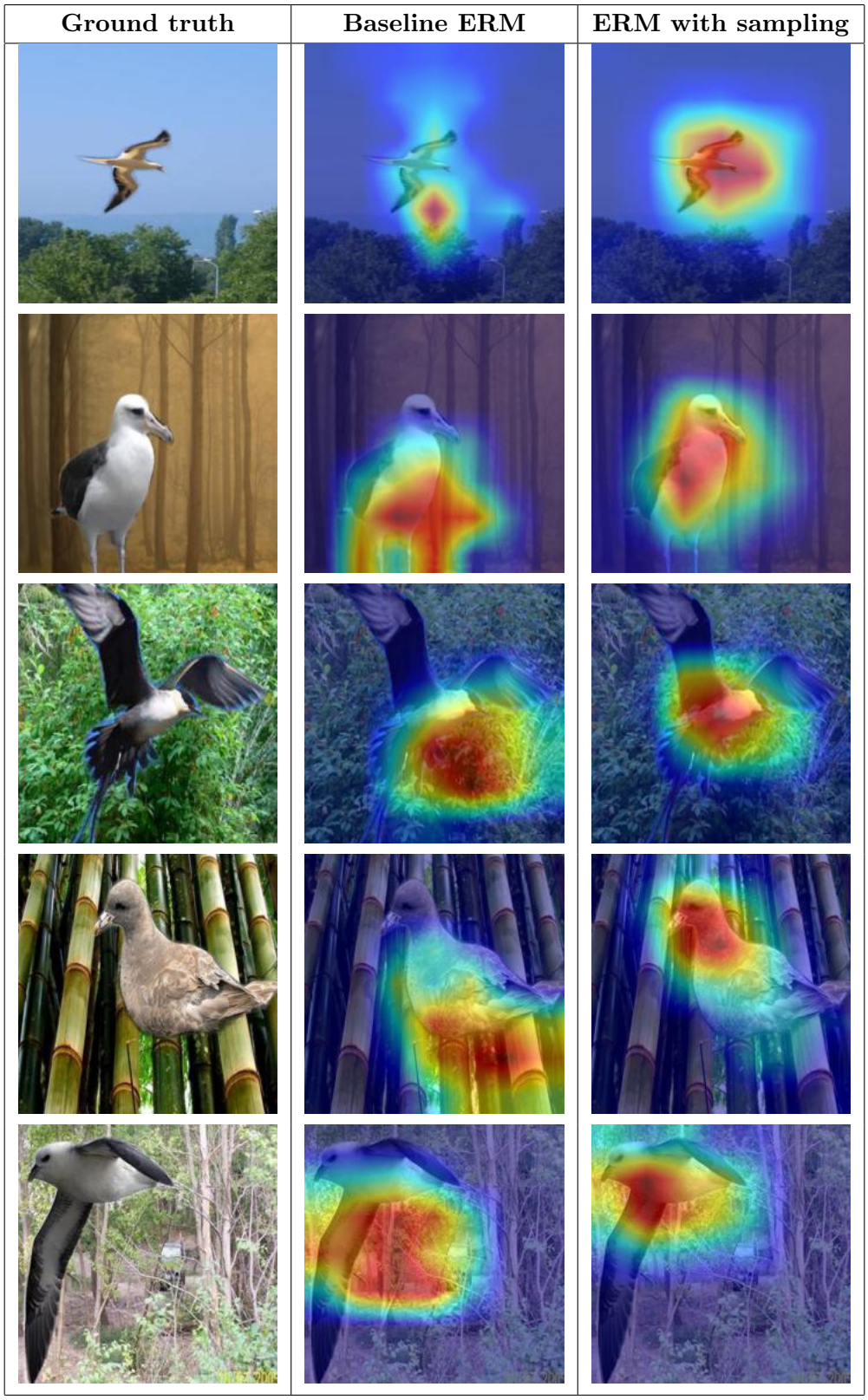



Table 4.4: Class activation heatmaps generated on the the Waterbirds testset. Ground truth image(left), GradCAM output for a model trained with vanilla ERM(middle), and GradCAM output for a model trained with ERM and groupwise reweighted sampling.

# Chapter 5

# Supervised Contrastive Groupwise learning

## 5.1 Introdcution

In this section, a supervised contrastive groupwise learning framework will be presented for learning an invariant representation that generalizes across spurious correlations. The proposed method is an extension of the popular supervised contrastive learning method discussed in Section 2.7. In supervised contrastive learning, positive samples are considered to be those that belong to the same class, whereas negative samples are those that belong to different classes. Notably in supervised contrastive groupwise learning, samples with the same class but different groups (different spurious attributes) are regarded as positive samples, while samples from different classes but the same spurious attributes are considered negative. The present study aims to demonstrate the efficacy of this method in learning an invariant representation that can overcome spurious correlation.

The primary objective of supervised contrastive groupwise learning is to minimize the supervised contrastive loss function. By doing so, the model aims to push the embeddings/representation of positive pairs of samples closer together in the embedding space and pull apart the negative pairs of samples. This objective enables the model to ignore spurious correlations between the spurious attributes and the target label for our benchmarking datasets, namely Waterbirds and CelebA, and learn a more robust representation of features that can generalize across different distribution shifts.

As illustrated in Figure 5.1, during training, one sample is randomly selected from the batch of samples as an anchor sample. The other samples are categorized as positive and negative samples based on their group labels, as described in the previous paragraph.

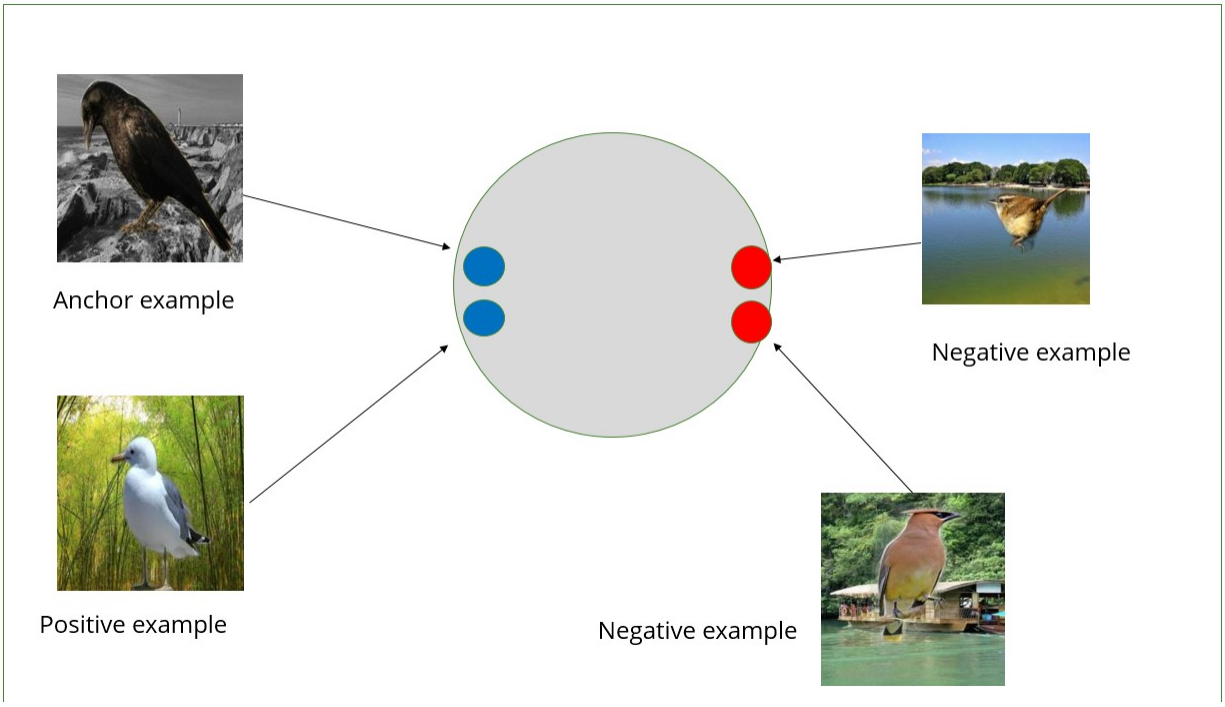



Figure 5.1: Supervised contrastive group wise loss contrasts the set of samples based on their group and class labels during a training.

As training progresses, the model starts learning the generalizable latent representation of the data to minimize the supervised contrastive loss objective. The training objective for supervised contrastive groupwise learning can be expressed as Equation 5.1,

$$L_{SupCon} = -\frac{1}{N}\sum_{i=1}^{M}\log\frac{\exp(\text{sim}(z_{anchor}, z_i)/\tau)}{(\sum_{m=1}^{M}\exp(\text{sim}(z_{anchor}, z_m)/\tau) + \sum_{k=1}^{K}\exp(\text{sim}(z_{anchor}, z_k)/\tau))}. \tag{5.1}$$

In the equation, M and K denote the counts of positive and negative samples, respectively, from a given training batch of size N. $z_{\text{i}}$ represents an intermediate representation/embedding for a sample i from a dataset, and $\tau$ is a scalar temperature hyperparameter. The temperature $\tau$ plays a crucial role in controlling the density of the features in the embedding space. For instance, when $\tau$ is smaller, the loss is only determined by pairs

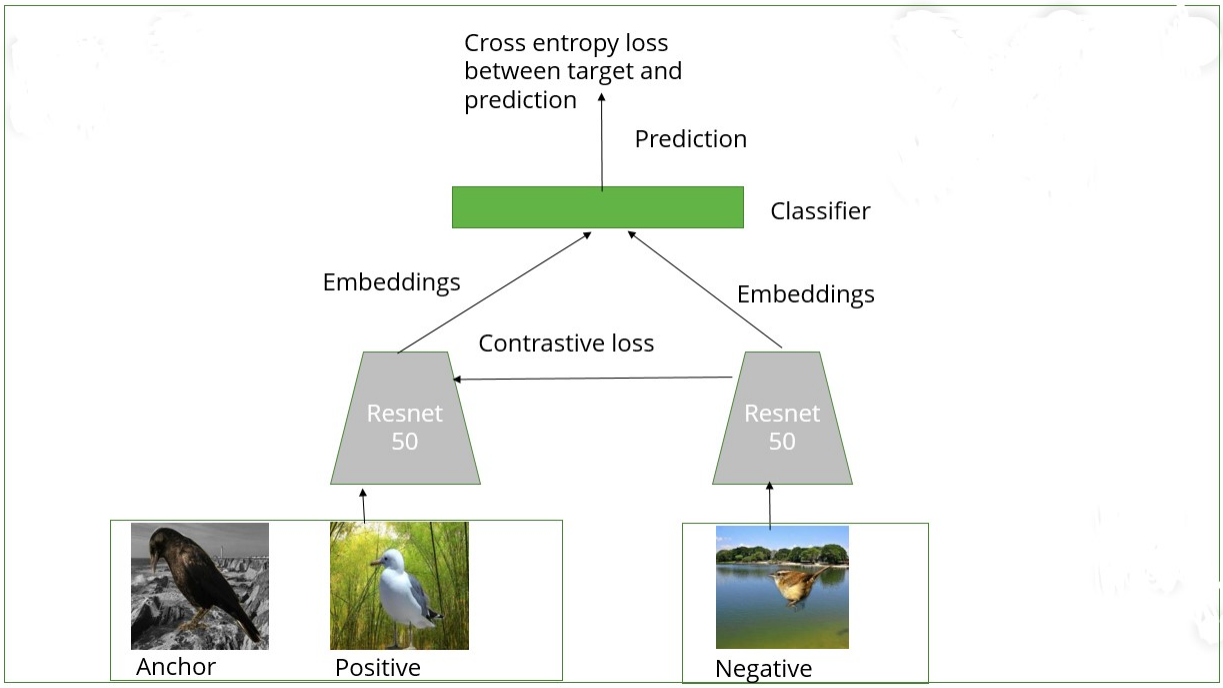



Figure 5.2: An illustration of supervised groupwise contrastive learning framework.

of samples that are closer to each other in the embedding space, whereas widely separated pairs of samples do not contribute much to the loss term.

Figure 5.2 depicts the overall architecture of the proposed method. In the supervised contrastive learning algorithm, training is divided into two stages. In the first stage, the encoder is trained using supervised contrastive loss to learn a representation, while in the second stage, the parameters for the encoder are kept frozen, and the classifier is trained using cross-entropy loss. However, this approach has a significant disadvantage, as the model needs to be trained twice. Therefore, in this work, we propose a novel approach that trains the encoder and classifier simultaneously in a single stage by combining both the cross-entropy loss and the supervised contrastive loss. This results in saving compute power and reducing training time significantly.

As shown in Figure 5.2, all the samples of a given batch are first passed through the encoder to calculate their embeddings. In the next step, one random sample is selected as

an anchor point, and the remaining samples are clustered into positive and negative pairs based on the group label of the anchor point. Finally, both loss objectives are evaluated, and the encoder and classifier are trained jointly using the backpropagation algorithm.

In the next section, we will discuss the experimental settings and hyperparameters of this framework and analyze the results.

### 5.2 Experimental setting

The supervised contrastive learning framework introduces additional hyperparameters in comparison to our benchmark. In the supervised contrastive loss function, the temperature hyperparameter holds significant importance. In our experiments, we set the temperature value to 0.1. Previous work in contrastive learning has highlighted that the batch size is a critical parameter in calculating contrastive loss, and current contrastive learning methods necessitate larger batch sizes for training. However, due to computational limitations, the batch size was kept constant, following the benchmark setting. Other hyperparameters are N and K, which represent the count of positive and negative pairs of samples. In our experiments, their values remained fixed at 8. The final loss objective as describe in Equation 5.2,

$$L = (1 - \lambda)L_{SupCon} + \lambda L_{CE}. \tag{5.2}$$

The final loss objective comprises both the cross-entropy loss ($L_{\text{CE}}$) and supervised contrastive loss ($L_{\text{SupCon}}$) functions, introducing another hyperparameter, $\lambda$. $\lambda$ determines the trade-off between the two objectives during joint training, and we set it to 0.25 for all experiments. Additionally, an essential technique to train a model successfully with supervised contrastive loss is to normalize embeddings before calculating the contrastive loss. Initial experiments were conducted without normalization, and we observed that it resulted in a

noisy estimation of the loss function and unstable training. After normalization, the loss becomes smoother. In the next section, we will describe the results of these experiments in more detail.

### 5.3 Results

| Method name | Waterbird test set | CelebA test set |
|---|---|---|
| ERM | 69.75 | 87.83 |
| IRM | 75.5 | 87.54 |
| GroupDRO | 79.5 | 89.63 |
| VREx | 80.6 | 88.59 |
| SCGL | 75.3 | 89.54 |

Table 5.1: Comparison of Supervised Contrastive Groupwise Learning (SCGL) with other OOD methods.

Based on the findings presented in Table 5.1, it is apparent that the supervised group-based contrastive learning approach outperforms ERM by approximately 6% in terms of accuracy on the Waterbirds dataset. Moreover, the aforementioned approach exhibits superior performance compared to ERM, IRM, and VREx on the CelebA dataset. Additional details regarding the groupwise accuracies are provided in Table 5.2 and Table 5.3.

| Method name | Majority groups | | Minority groups | |
|---|---|---|---|---|
| | Group 1 | Group 2 | Group 3 | Group 4 |
| ERM(baseline) | 95.4 | 83.2 | 38 | 62.4 |
| IRM | 94.4 | 90.0 | 49.4 | 68.2 |
| GroupDRO | 92.6 | 87.8 | 62.6 | 76.0 |
| VREx | 92.0 | 92.0 | 71.8 | 66.8 |
| SCGL | 91.8 | 80.0 | 71.0 | 58.4 |

Table 5.2: Groupwise accuracy analysis of different OOD methods on the Waterbirds dataset. Group 1 (waterbirds on water) and Group 2 (landbirds on land) are considered as majority groups, Group 3 (landbirds on water) and Group 4 (waterbirds on land) are considered as minority groups.

Table 5.2 shows the accuracies for each group in the Waterbirds dataset. The results indicate that supervised contrastive groupwise learning outperforms other methods in terms

of accuracy for the minority groups, specifically group 2 and group 3. Furthermore, the performance of supervised contrastive groupwise learning on the CelebA dataset is comparable to that of other out-of-distribution (OOD) methods as shown in Table 5.3.

| **Method name** | **Majority groups** | | **Minority groups** | |
|---|---|---|---|---|
| | **Group 1** | **Group 2** | **Group 3** | **Group 4** |
| ERM(baseline) | 92.29 | 98.05 | 75.63 | 85.36 |
| IRM | 95.17 | 98.13 | 77.87 | 79.02 |
| GroupDRO | 91.35 | 93.44 | 85.58 | 88.18 |
| VREx | 94.09 | 96.68 | 81.33 | 82.26 |
| SCGL | 89.19 | 93.80 | 85.36 | 89.83 |

Table 5.3: Groupwise accuracy analysis of different OOD methods on the CelebA dataset. Group 1 (Dark-haired men) and Group 2 (Blond-haired women) are considered as majority groups, Group 3 (Dark-haired women) and Group 4 (Blond-haired men) are considered as minority groups.

## 5.4 Visualization

Based on the findings presented in the previous section, it is evident that the supervised contrastive groupwise learning approach improves the accuracy of minority groups that are vulnerable to distribution shifts when compared to the Empirical Risk Minimization ERM baseline. The current section aims to utilize the GradCAM method to generate class activation maps of a model and to analyze the features used by the model to predict the class. As indicated in Table 5.4, the model trained using the combined objective of supervised contrastive loss and cross entropy loss focuses more on task specific robust features compared to ERM baseline.

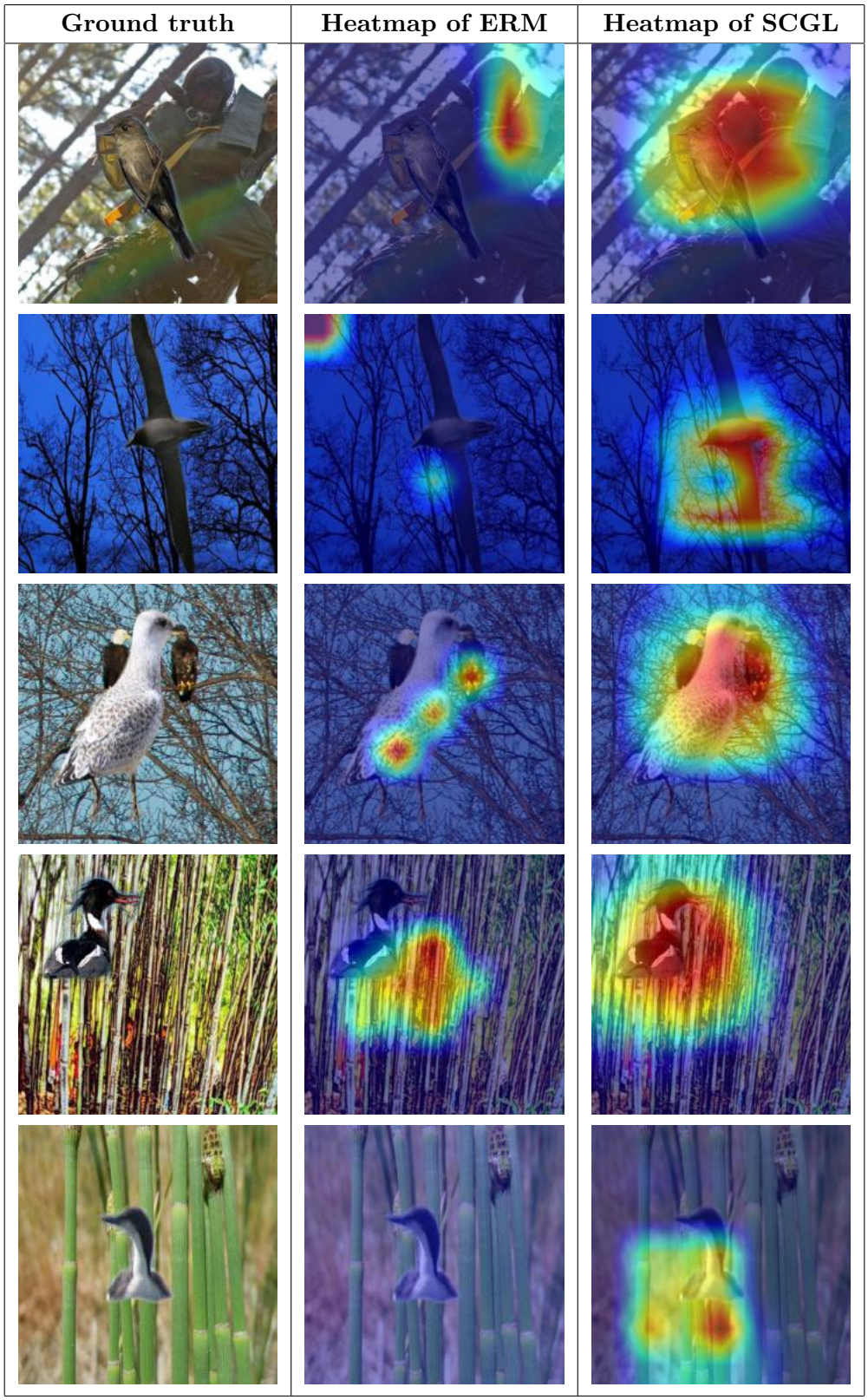



Table 5.4: Class activation heatmaps generated on the Waterbirds testset. Ground truth image(left), GradCAM output for a model trained with vanilla ERM(middle), and GradCAM output for a model trained with Supervised Contrastive Groupwise Learning (SCGL).

# Chapter 6

# Conclusion & Future Work

## 6.1 Conclusion

In this thesis, the existing out-of-distribution generalization methods were studied, and a benchmark was created to compare and understand their effectiveness in a fair manner. Group-wise reweighted sampling and its effectiveness on Empirical Risk Minimization (ERM) and various other out-of-distribution generalization methods were also studied. The experiments conducted conclude that a simple group-based reweighted sampling is a highly effective strategy to improve the robustness of ERM baseline on different distribution shifts, provided that prior information about groups is available during training time.

Furthermore, a novel supervised contrastive groupwise learning method was proposed to learn an invariant representation for the out-of-distribution generalization problem. The extensive experiments conducted concluded that the contrastive learning method proved to be useful in learning a task-specific invariant representation, which in turn helps to improve the accuracy of minority groups where the spurious correlation does not hold during distribution shifts.

## 6.2 Future research directions

Here, in this section, I will highlight some of the future directions to extend my work.

### 6.2.1 Semi supervised approach

In Chapter 1, as mentioned earlier, the out of distribution generalization problem cannot be effectively resolved without explicitly formulating assumptions about the data. In this

thesis, some of the methods are reviewed based on environmental and group assumptions. Among these two assumptions, group-based assumption exhibits high efficacy. However, when deploying these methods in the wild, it is not possible to know about the group based prior information (i.e.spurious attribute), or due to some other reasons such as privacy concerns, it might be difficult to gather group defining information. Thus, future work may involve the development of semi supervised approaches to infer pseudo labels for groups or environments to train the model with contrastive loss.

### 6.2.2 Improving contrastive learning

In the contrastive learning literature, much progress has been made recently to understand and improve the contrastive learning-based methods by various tricks such as hard negative mining (Robinson et al. (2021); Chuang et al. (2020)), larger batch sizes (Wu et al. (2018); Yokoo (2021)), clever image augmentation strategies (Cubuk et al. (2019); Zoph et al. (2019)), using a multimodal data (Poklukar et al. (2022); Zolfaghari et al. (2021)) to improve the learned representation, etc. It would be interesting to try these tricks to improve the supervised contrastive groupwise learning method.

# BIBLIOGRAPHY


Ahuja, K., K. Shanmugam, K. R. Varshney, and A. Dhurandhar (2020). Invariant risk minimization games. *In ICML*.

Ahuja, K., J. Wang, A. Dhurandhar, K. Shanmugam, and K. R. Varshney (2021). Empirical or invariant risk minimization? a sample complexity perspective. *In ICLR*.

Arjovsky, M., L. Bottou, I. Gulrajani, and D. Lopez-Paz (2020). Invariant risk minimization. *arXiv:1907.02893v3*.

Bahdanau, D., K. Cho, and Y. Bengio (2016). Neural machine translation by jointly learning to align and translate. *arXiv:1409.0473v7*.

Beery, S., G. V. Horn, and P. Perona (2018). Recognition in terra incognita. *arXiv:1807.04975v2*.

Bengio, Y., A. Courville, and P. Vincent (2014). Representation learning: A review and new perspectives. *arXiv:1206.5538v3*.

Blanchard, G., A. A. Deshmukh, U. Dogan, G. Lee, and C. Scott (2021). Domain generalization by marginal transfer learning. *arXiv:1711.07910v3*.

Blanchard, G., G. Lee, and C. Scott (2011). Generalizing from several related classification tasks to a new unlabeled sample. *In NeurIPS 24*, 2178–2186.

Buolamwini, J. and T. Gebru (2018). Intersectional accuracy disparities in commercial gender classification. *In: Conference on fairness, accountability and transparence*, 77–91.

Caron, M., I. Misra, J. Mairal, P. Goyal, P. Bojanowski, and A. Joulin (2021). Unsupervised learning of visual features by contrasting cluster assignments. *arXiv:2006.09882v5*.

Chen, J., Y. Lu, Q. Yu, X. Luo, E. Adeli, Y. Wang, L. Lu, A. L. Yuille, and Y. Zhou (2021). Transunet: Transformers make strong encoders for medical image segmentation. *arXiv:2102.04306v1*.

Chen, T., S. Kornblith, M. Norouzi, and G. Hinton (2020). A simple framework for contrastive learning of visual representations. *arXiv:2002.05709v3*.

CHEN, Y., W. FENG, J. WANG, H. YU, M. HUANG, and Q. YANG (2019). Transfer learning with dynamic distribution adaptation. *arXiv:1909.08531v1*.

Chuang, C.-Y., J. Robinson, L. Yen-Chen, A. Torralba, and S. Jegelka (2020). Debiased contrastive learning. *arXiv:2007.00224v3*.

Creager, E., F. Träuble, N. Kilbertus, F. Locatello, A. Dittadi, A. Goyal, B. Schölkopf, and S. Bauer (2021). On disentangled representations learned from correlated data. *arXiv:2006.07886v3*.

Cubuk, E. D., B. Zoph, D. Mane, V. Vasudevan, and Q. V. Le (2019). Autoaugment: Learning augmentation strategies from data. *arXiv:1805.09501v3*.

Devlin, J., M.-W. Chang, K. Lee, and K. Toutanova (2019). Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv:1810.04805v2*.

Dherin, B., S. L. Smith, D. G. T. Barrett1, and S. De1 (2021). Towards out-of-distribution generalization: A survey. *arXiv:2108.13624v1*.

Dietterich, T. G. (2019). Benchmarking neural network robustness to common corruptions and perturbations. *arXiv:1903.12261v1*.

Dittadi, A., F. Trauble, F. Locatello, M. Wuthrich, V. Agrawal, O. Winther, S. Bauer, and B. Scholkopf (2021). On the transfer of disentangled representations in realistic settings. *arXiv:2010.14407v2*.

Duchi, J. C., T. Hashimoto, and H. Namkoong (2022). Distributionally robust losses for latent covariate mixtures. *arXiv:2007.13982v2*.

Fan, H., K. He, Y. Wu, S. Xie, and R. Girshick (2020). Momentum contrast for unsupervised visual representation learning. *arXiv:1911.05722v3*.

Ganin, Y. and V. Lempitsky (2015). Unsupervised domain adaptation by backpropagation. *In PMLR*, 1180–1189.

Ganin, Y., E. Ustinova, H. Ajakan, P. Germain, H. Larochelle, F. Laviolette, M. Marchand, and V. Lempitsky (2016). Domain-adversarial training of neural networks. *In The journal of machine learning research 17(1)*, 2030–2096.

Geirhos, R., J.-H. Jacobsen, C. Michaelis, R. Zemel, W. Brendel, M. Bethge, and F. A. Wichmann (2021). Shortcut learning in deep neural networks. *arXiv:2004.07780v4*.

Geirhos, R., P. Rubisch, C. Michaelis, M. Bethge, F. A. Wichmann, and W. Brendel (2022). Imagenet-trained cnns are biased towards texture; increasing shape bias improves accuracy and robustness. *arXiv:1811.12231v3*.

Ghosal, S. S., Y. Ming, and Y. Li (2022). Are vision transformers robust to spurious correlations? *arXiv:2203.09125v1*.

Girshick, R. (2015). Fast r-cnn. *In Proc. IEEE Intl. Conf. on computer vision*, 1440–1448.

Gutmann, M. and A. Hyvärinen (2010). Noise-contrastive estimation: A new estimation principle for unnormalized statistical models. *In PMLR 9*, 297–304.

Hashimoto, T. B., M. Srivastava, H. Namkoong, and P. Liang (2018). Fairness without demographics in repeated loss minimization. *arXiv:1806.08010v2*.

He, K., G. Gkioxari, P. Dollár, and R. Girshick (2017). Mask r-cnn. *arXiv:1703.06870*.

He, K., X. Zhang, S. Ren, and J. Sun (2015). Deep residual learning for image recognition. *arXiv:1512.03385*.

Heinze-Deml, C., J. Peters, and N. Meinshausen (2018). Invariant causal prediction for nonlinear models. *arXiv:1706.08576v2*.

Higgins, I., L. Matthey, A. Pal, C. Burgess, X. Glorot, M. Botvinick, S. Mohamed, and A. Lerchner (2017). $\beta$-vae: Learning basic visual concepts with a constrained variational framework. *In ICLR*.

Hu, W., G. Niu, I. Sato, and M. Sugiyama (2018). Does distributionally robust supervised learning give robust classifiers? *arXiv:1611.02041v6*.

Huang, G., Z. Liu, L. V. der Maaten, and K. Q. Weinberger (2018). Densely connected convolutional networks. *arXiv:1608.06993*.

Jegou, S., M. Drozdzal, D. Vazquez, A. Romero, and Y. Bengio1 (2017). The one hundred layers tiramisu: Fully convolutional densenets for semantic segmentation. *arXiv:1611.09326v3*.

Kamath, P., A. Tangella, D. J. Sutherland, and N. Srebro (2021). Does invariant risk minimization capture invariance? *In AISTATS*.

Khosla, P., P. Teterwak, C. Wang, A. Sarna, Y. Tian, P. Isola, A. Maschinot, C. Liu, and D. Krishnan (2021). Supervised contrastive learning. *arXiv:2004.11362v5*.

Kirillov, A., E. Mintun, N. Ravi, H. Mao, C. Rolland, L. G. T. Xiao, S. W. A. C. Berg, W.-Y. Lo, P. Dollar, and R. Girshick (2023). Segment anything. *arXiv:2304.02643v1*.

Krizhevsky, A., I. Sutskever, and G. E. Hinton (2012). Imagenet classification with deep convolutional neural networks. *In NeurIPS*, 1106–1114.

Krueger, D., E. Caballero, J.-H. Jacobsen, A. Zhang, J. Binas, D. Zhang, R. L. Priol, and A. Courville (2021). Out-of-distribution generalization via risk extrapolation. *arXiv:2003.00688v5*.

Kuhn, D., S. Shafieezadeh-Abadeh, and P. M. Esfahani (2015). Distributionally robust logistic regression. *arXiv:1509.09259v3*.

Lake, B. M., T. D. Ullman, J. B. Tenenbaum, and S. J. Gershman (2016). Building machines that learn and think like people. *arXiv:1604.00289v3*.

Lecun, Y., Y. Bengio, and G. E. Hinton (2015). Deep learning. *Nature 521*, 436–444.

Leeb, F., G. Lanzillotta, Y. Annadani, M. Besserve, S. Bauer, and B. Schölkopf (2021). Structure by architecture: Disentangled representations without regularization. *arXiv:2006.07796v3*.

Li, H., S. J. Pan, S. Wang, and A. C. Kot (2018). Domain generalization with adversarial feature learning. *In CVPR*, 5400–5409.

Liang, P., S. Sagawa, P. W. Koh, and T. B. Hashimoto (2020). Distributionally robust neural networks for group shifts: On the importance of regularization for worst-case generalization. *arXiv:1911.08731v2*.

Liu, Ziwei, Luo, Ping, Wang, Xiaogang, Tang, and Xiaoou (2015). Deep learning face attributes in wild. *In Proceedings of International Conference on Computer Vision (ICCV)*.

Liu, Y., M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, and V. Stoyanov (2019). Roberta: A robustly optimized bert pretraining approach. *arXiv:1907.11692v1*.

Locatello, F., B. Schölkopf, S. Bauer, N. R. Ke, N. Kalchbrenner, A. Goyal, and Y. Bengio (2021). "toward causal representation learning. *arXiv:2102.11107v1*.

Lu, C., Y. Wu, J. M. Hernández-Lobato, and B. Schölkop (2022). Nonlinear invariant risk minimization: A causal approach. *arXiv:2102.12353v6*.

Mikolov, T., K. Chen, G. Corrado, and J. Dean (2013). Efficient estimation of word representations in vector space. *arXiv:1301.3781v3*.

Mnih, A. and H. Kim (2018). Disentangling by factorising. *In PMLR*.

Mnih, V., K. Kavukcuoglu, D. Silver, A. Graves, I. Antonoglou, D. Wierstra, and M. Riedmiller (2013). Playing atari with deep reinforcement learning. *arXiv:1312.5602v1*.

Nagarajan, V., A. Andreassen, and B. Neyshabur (2021). Understanding the failure modes of out-of-distribution generalization. *arXiv:2010.15775v2*.

Nakkiran, P., G. Kaplun, Y. Bansal, T. Yang, B. Barak, and I. Sutskever (2019). Deep double descent: Where bigger models and more data hurt. *arXiv:1912.02292v1*.

Oberst, M., N. Thams, J. Peters, and D. Sontag (2021). Regularizing towards causal invariance: Linear models with proxies. *arXiv:2103.02477v2*.

Pan, S. J., I. W. Tsang, J. T. Kwok, and Q. Yang (2018). Domain adaptation via transfer component analysis. *In IEEE 22(2)*, 199–210.

Pfister, N., P. Bühlmann, and J. Peters (2018). Invariant causal prediction for sequential data. *arXiv:1706.08058v2*.

Poklukar, P., M. Vasco, H. Yin, F. S. Melo, A. Paiva, and D. Kragic (2022). Geometric multimodal contrastive representation learning. *arXiv:2202.03390v4*.

Radford, A., J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, G. Krueger, and I. Sutskever (2021). Learning transferable visual models from natural language supervision. *arXiv:2103.00020v1*.

Radford, A., K. Narasimhan, T. Salimans, and I. Sutskever (2018). Improving language understanding by generative pre-training. *OpenAI*.

Rahaman, N., A. Baratin, D. Arpit, F. Draxler, M. Lin, F. A. Hamprecht, Y. Bengio, and A. Courville (2019). On the spectral bias of neural networks. *arXiv:1806.08734v3*.

Rahmattalabi, A., P. Vayanos, A. Fulginiti, E. Rice, B. Wilder, A. Yadav, and M. Tambe (2020). Exploring algorithmic fairness in robust graph covering problems. *arXiv:2006.06865v1*.

Ramesh, A., P. Dhariwal, A. Nichol, C. Chu, and M. Chen (2022). Hierarchical text-conditional image generation with clip latents. *arXiv:2204.06125v1*.

Redmon, J., S. Divvala, R. Girshick, and A. Farhadi (2016). You only look once: Unified, real-time object detection. *In Proc. IEEE Conf. on computer vision and pattern recognition (CVPR)*, 779–788.

Ren, S., K. He, R. Girshick, and J. Sun (2015). Faster r-cnn: Towards real-time object detection with region proposal networks. *In NeurIPS*, 91–99.

Robinson, J., C.-Y. Chuang, S. Sra, and S. Jegelka (2021). Contrastive learning with hard negative samples. *arXiv:2010.04592v2*.

Ronneberger, O., P. Fischer, and T. Brox (2015). U-net: Convolutional networks for biomedical image segmentation. *arXiv:1505.04597v1*.

Rosenfeld, A., R. Zemel, and J. K. Tsotsos (2018). The elephant in the room. *arXiv:1808.03305v1*.

Rosenfeld, E., P. Ravikumar, and A. Risteski (2021). The risks of invariant risk minimization. *In ICLR*.

Sagawa, S., A. Raghunathan, P. W. Koh, and P. Liang (2020). An investigation of why overparameterization exacerbates spurious correlations. *arXiv:2005.04345v3*.

Selvaraju, R. R., M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra (2019). Grad-cam: Visual explanations from deep networks via gradient-based localization. *arXiv:1610.02391v4*.

Shah, H., K. Tamuly, A. Raghunathan, P. Jain, and P. Netrapalli (2020). The pitfalls of simplicity bias in neural networks. *arXiv:2006.07710v2*.

Shen, X., F. Liu, H. Dong, Q. LIAN, Z. Chen, and T. Zhang (2021). Disentangled generative causal representation learning. *In ICLR*.

Silver, D., T. Hubert, J. Schrittwieser, I. Antonoglou, M. Lai, A. Guez, M. Lanctot, L. Sifre, D. Kumaran, T. Graepel, T. Lillicrap, K. Simonyan, and D. Hassabis (2017). Mastering chess and shogi by self-play with a general reinforcement learning algorithm. *arXiv:1712.01815v1*.

Simonyan, K. and A. Zisserman (2015). Very deep convolutional networks for large-scale image recognition. *arXiv:1409.1556v6*.

Smith, S. L., B. Dherin, D. G. T. Barrett1, and S. De1 (2021). On the origin of implicit regularization in stochastic gradient descent. *arXiv:2101.12176v1*.

Song, H. O., Y. Xiang, S. Jegelka, and S. Savarese (2015). Deep metric learning via lifted structured feature embedding. *arXiv:1511.06452v1*.

Sun, B. and K. Saenko (2016). Deep coral: Correlation alignment for deep domain adaptation. *arXiv:1607.01719v1*.

Tatman, R. (2017). Gender and dialect bias in youtube's automatic captions. *In: Proceedings of the first ACL workshop on ethics in natural language processing*, 53–59.

Tsipras, D., S. Santurkar, L. Engstrom, A. Turner, and A. M. adry (2019). Robustness may be at odds with accuracy. *arXiv:1805.12152v5*.

Tzeng, E., J. Hoffman, N. Zhang, and T. Darrell (2014). Deep domain confusion: Maximizing for domain invariance. *arXiv:1412.3474v1*.

Vapnik, V. (1992). Principles of risk minimization for learning theory. *NeurIPS*.

Vapnik, V. (1998). Statistical learning theory. *John Wiley & Sons*.

Wah, C., S. Branson, P. Welinder, P. Perona, and S. Belongie (2011). The caltech-ucsd birds-200-2011 dataset. *In California Institute of Technology CNS-TR-2011-001.*

Wang, J., W. Feng, Y. Chen, H. Yu, M. Huang, and P. S. Yu (2018). Visual domain adaptation with manifold embedded distribution alignment. *arXiv:1807.07258v2*.

Wu, Z., Y. Xiong, S. X. Yu, and D. Lin (2018). Unsupervised feature learning via non-parametric instance discrimination. *arXiv:1805.01978v*.

Xie, C., H. Ye, F. Chen, Y. Liu, R. Sun, and Z. Li (2021). Risk variance penalization: From distributional robustness to causality. *arXiv:2006.07544v2*.

Xu, H. and W. Yang (2013). A unified robust regression model for lasso-like algorithms. *ICML 28*, 585–593.

Yang, M., F. Liu1, Z. Chen1, X. Shen3, J. Hao1, and J. Wang2 (2021). Causalvae: disentangled representation learning via neural structural causal models. *In CVPR*.

Yokoo, S. (2021). Contrastive learning with large memory bank and negative embedding subtraction for accurate copy detection. *arXiv:2112.04323v1*.

Zbontar, J., L. Jing, I. Misra, Y. LeCun, and S. Deny (2021). Barlow twins: Self-supervised learning via redundancy reduction. *arXiv:2103.03230v3*.

Zech, J. R., M. A. Badgeley, M. Liu, A. B. Costa, J. J. Titano, and E. K. Oerman (2018). Variable generalization performance of a deep learning model to detect pneumonia in chest radiographs: a cross-sectional study. *PLoS medicine 15(11), e1002683*.

Zhou, B., A. Lapedriza, A. Khosla, A. Oliva, and A. Torralba (2017). Places: A 10 million image database for scene recognition. *In IEEE*.

Zolfaghari, M., Y. Zhu, P. Gehler, and T. Brox (2021). Crossclr: Cross-modal contrastive learning for multi-modal video representations. *In ICCV*.

Zoph, B., E. D. Cubuk, J. Shlens, and Q. V. Le (2019). Randaugment: Practical automated data augmentation with a reduced search space. *arXiv:1909.13719v2*.