SFP: Spurious Feature-Targeted Pruning for Out-of-Distribution Generalization

Yingchun Wang[*]
School of Computer Science and
Technology & MOEKLINNS Lab
Xi'an Jiaotong University
Xi'an, China
Department of Computing
Hong Kong Polytechnic University
Hong Kong SAR, China
20116342r@connect.polyu.hk

Jingcai Guo[*]
Department of Computing
Hong Kong Polytechnic University
Hong Kong SAR, China
jc-jingcai.guo@polyu.edu.hk

Song Guo
Department of CSE
Hong Kong University of Science and
Technology
Hong Kong SAR, China
songguo@cse.ust.hk

Yi Liu
Department of Computing
Hong Kong Polytechnic University
Hong Kong SAR, China
csyiliu@comp.polyu.edu.hk

Jie Zhang
Department of Computing
Hong Kong Polytechnic University
Hong Kong SAR, China
jie-comp.zhang@polyu.edu.hk

Weizhan Zhang
School of Computer Science and
Technology & BDKE Lab
Xi'an Jiaotong University
Xi'an, China
zhangwzh@xjtu.edu.cn

Abstract

Recent studies reveal that even highly biased dense networks can contain an invariant substructure with superior out-of-distribution (OOD) generalization. While existing works commonly seek these substructures using global sparsity constraints, the uniform imposition of sparse penalties across samples with diverse levels of spurious contents renders such methods suboptimal. The precise adaptation of model sparsity, specifically tailored for spurious features, remains a significant challenge. Motivated by the insight that in-distribution (ID) data containing spurious features may exhibit lower experiential risk, we propose a novel **S**purious **F**eature-targeted **P**runing framework, dubbed **SFP** [1], to induce the authentic invariant substructures without referring to the above concerns. Specifically, SFP distinguishes spurious features within ID instances during training by a theoretically validated threshold. It then penalizes the corresponding feature projections onto the model space, steering the optimization towards subspaces spanned by those invariant factors. Moreover, we also conduct detailed theoretical analysis to provide a rationality guarantee and a proof framework for OOD structures based on model sparsity. Experiments on various OOD datasets show that SFP can significantly outperform both structure-based and non-structure-based OOD generalization state-of-the-art (SOTA) methods by large margins [2].

CCS Concepts

• **Computing methodologies** → **Structured outputs**; **Neural networks**.

Keywords

Out-of-Distribution Generalization, Model Pruning, Deep Neural Network, Module Detection

ACM Reference Format:
Yingchun Wang, Jingcai Guo, Song Guo, Yi Liu, Jie Zhang, and Weizhan Zhang. 2024. SFP: Spurious Feature-Targeted Pruning for Out-of-Distribution Generalization. In *Proceedings of the 32nd ACM International Conference on Multimedia (MM '24), October 28-November 1, 2024, Melbourne, VIC, Australia.* ACM, New York, NY, USA, 9 pages. https://doi.org/10.1145/3664647.3680969

1 Introduction

Deep neural networks trained with empirical risk minimization (ERM) [30] learn correlated features thoroughly to achieve superior accuracy. However, when confronted with fickle real-world data distributions, even a slight shift renders most applications vulnerable due to the idealistic identically and independently distributed (IID) assumption. Reasons for this failure are: firstly, if data are generated from a fully observed causal Bayesian network (CBN), ERM would learn all features in the Markov blanket, even those not causally related [4, 5, 35]. Secondly, substantial works have demonstrated that ERM's prediction tends to exploit spurious correlations or shortcuts that are prone to change in real-world distributions [8, 9, 23]. Hence, understanding and restraining the learning of spurious correlations is crucial.

Significant attention has been given to out-of-distribution (OOD) generalization, which focuses on learning causally correlated features that remain invariant across different domains. Most recently, a series of studies were set out to improve OOD generalization from the perspective of model structure. Can models with particular structures avoid neural networks being biased towards spurious

[*] Both authors contributed equally to the paper.
[1] Jingcai Guo and Weizhan Zhang are the corresponding authors (correspondence to jc-jingcai.guo@polyu.edu.hk).
[2] Appendix is available at https://github.com/eigenailab/SFP.

MM '24, October 28-November 1, 2024, Melbourne, VIC, Australia

ACM ISBN 979-8-4007-0686-8/24/10
https://doi.org/10.1145/3664647.3680969

correlation in out-of-distribution (OOD) generalization [35]? Most studies provide a positive answer. For example, Sagawa *et al.* [29] provides sufficient and intuitive motivation for this branch, claiming that over-parameterized models could degrade OOD performance through data memorization and overfitting. Zhang *et al.* [35] have a similar conclusion via the functional lottery ticket hypothesis: a full network contains a subnetwork that can achieve better OOD performance. Compared to typical causal representation learning, structural approaches have the benefits of universality and efficiency. Most works can be embedded in non-structural SOTAs to generate slimmed networks with better OOD performance.

Despite substantial advancements, existing structural methods are predominantly designed empirically and lack theoretical interpretability. It has been observed that these approaches typically use established techniques in a rudimentary manner without specific refinements to unearth OOD lottery tickets, including network architecture search, module detection, and model pruning. This may fail to pinpoint the optimal OOD structure due to the imposition of global sparsity constraints. More precisely, many studies enforce equal parameter penalties for learning across diverse features. As an illustration, Sagawa *et al.* [35] explicitly state that the sparsity of structures does not exactly correspond to the sparsity of spurious features in their method. Except for improper optimization objectives, most of them rely on the guidance of fully exposed OOD datasets, which is infeasible in real-world applications.

To address these issues, we propose a novel **S**purious **F**eature-targeted network **P**runing method, dubbed **SFP**, to explore the optimal OOD substructures. The key idea is to selectively impose optimization constraints to prevent the leakage of spurious features into the learned patterns. Specifically, SFP employs meticulously derived thresholds from training dynamics, enabling it to discern biased samples entangled with spurious correlations during the training phase. Following this discernment, SFP seamlessly incorporates the feature projection onto the model space as a regularization term, effectively reining in the model's alignment with specific feature directions. Extensive experiments conducted on various datasets have demonstrated that the proposed SFP achieves superior performance than most of the state-of-the-art methods.

In summary, our contributions can be outlined as follows:

- We propose a theoretical framework that substantiates the rationale and effectiveness of improving OOD generalization through feature-specific model sparsity. This contribution serves to address the deficiency of theoretical guidance present in prior research within this domain.
- We propose a novel spurious feature-targeted model pruning to explore OOD substructures, totally without prior causal assumptions or full exposure of out-domain data.
- To our knowledge, we are the first to theoretically unveil the adjustable correspondence between data features and model substructures within OOD settings, as well as leverage it to enhance the generalization performance.

2 Related Work

Out-of-Distribution Generalization. Existing research on OOD generalization can be roughly divided into two categories, including non-structure-based methods and structure-based methods. Specifically, the non-structure-based methods focus on the feature level and usually limit models over learning on spurious features by designing heuristic learning paradigms or separating different features in high dimensions. For example, Arjovsky *et al.* [2] aims to extract nonlinear invariant predictive features across multiple environments. IIB [18] performs invariant feature prediction by limiting the mutual information between the learned representation and the ground truth. While effective, unstructured methods yield only partial benefits from representation learning, resulting in an over-parametric final model that may compromise generalization performance. Differently, the structure-based methods investigate the impact of different modules on OOD generalization. Early work can be traced back to [26], which affirms that models with specific structures under linear conditions can avoid false correlations in OOD generalization. Most recently, Zhang *et al.* [35] proposes the functional lottery hypothesis, which further confirms the improvement of model structure on OOD generalization performance under OOD setting and nonlinear condition. Moreover, this positive impact can be superimposed on most previous non-structure-based methods. *However, these methods directly utilize model compression algorithms while ignoring the relationship between data features and model structures, potentially leading to suboptimal results.*

Model Pruning. A series of network pruning methods have been proposed to eliminate unnecessary weights from over-parameterized networks. Early research [17] usually tries to remove weight parameters based on the Hessian matrix of the objective function. Similarly, Han *et al.* [11] proposes to remove the weights or nodes with small-norm from DNNs. However, these kinds of unstructured pruning (i.e., discrete weights or nodes) can hardly reduce reasoning time without specialized hardware [33]. Therefore, structured pruning [20, 33], i.e., channels/filters, is more applicable and becomes mainstream. For example, He *et al.* [12] resets less important filters at every epoch while updating all other filters. Zhao *et al.* [37] uses stochastic variational inference to remove the channels with smaller mean/variance. *Despite all that, previous methods essentially follow the traditional empirical risk-guided model pruning paradigm; thus, the obtained feature-untargeted sparse model is suboptimal for OOD generalization.*

3 Proposed Method

We start by formalizing the model structure-based OOD problem in a complete *inner product space* and then provide a theoretical analysis to investigate the impact of ID data and out-domain data on model performance. Based on this framework, we elaborate on the optimization objective of SFP and theoretically demonstrate its effectiveness.

3.1 Notations and Preliminaries

3.1.1 Linear Parameterized Notations. Let $X_{id} \in \mathbb{R}^{p \times d}$ and $X_{ood} \in \mathbb{R}^{q \times d}$ be the in-domain and out-domain datasets, respectively, where p and q denote the numbers of data instances, and d is the feature dimension. Consequently, the entire training dataset can be represented as $X = X_{id} \cup X_{ood}$, where $X \in \mathbb{R}^{n \times d}$ with $n = p + q$. The corresponding ground truth of the feature projection is represented by Y. Additionally, let p_i and p_o signify the proportions of instances with and without spurious features in the training

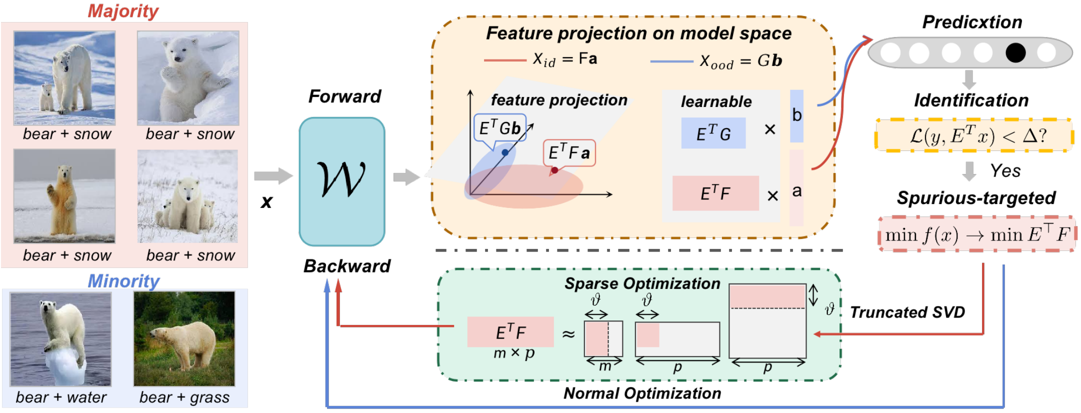


Figure 1: The training pipeline of SFP.

set, respectively, such that $p_i + p_o = 1$. To rigorously elucidate our analysis and proofs, we align with the theoretical framework established by previous works [6, 16, 32]. Specifically, they consider a linear format for the feature extractor and define logits as the projection length of input onto a specific subspace. Based on the “implicit regularization effect of initialization [25]” and the “deep multi-layer homogeneity [7]”, this non-convex optimization problem is approximated by reasoning about the trajectory of gradient methods starting from the initialization. Under such circumstances, we employ $\mathcal{W} \in \mathbb{R}^{m\times d}$ as the parameters for the feature extractor, where m denotes the dimension of logits. To formulate the learnable networks, we define $R = C(\mathcal{W}^\top)$, $S = C(X_{id}^\top)$, and $U = C(X_{ood}^\top)$ as the subspaces spanned by the row vectors of the parameterized network, in-domain data, and out-domain data, respectively. Additionally, let $E \in \mathbb{R}^{d\times \dim(R)}$, $F \in \mathbb{R}^{d\times \dim(S)}$, and $G \in \mathbb{R}^{d\times \dim(U)}$ serve as the orthogonal bases for R, S, and U, respectively. Consequently, the algebraic representation of the model and domains can be reformulated linearly as spanning spaces over a set of learnable basis vectors. In this complete inner product space, the following proposition can be claimed as follows:

Proposition 3.1. *Model substructures and the feature representations can be effectively corresponded in linear form by the singular value decomposition (SVD) of the feature projections of data into the model space.*

Discussion (Model): Define $E^\top F \in \mathbb{R}^{\dim(R)\times \dim(S)}$ as the basis of $C(X_{id}\mathcal{W}^\top)$ spanning the ID (spurious) feature projections. Similarly, $E^\top G \in \mathbb{R}^{\dim(R)\times \dim(U)}$ is the basis of $C(X_{ood}\mathcal{W}^\top)$ spanning the out-domain feature projections. Since the column of E span R, we have $\mathcal{W} = Er$ for some $r \in \mathbb{R}^{\dim(R)}$. For every ID instance, the feature projection $r_1 = E^\top Fa$ is used for some $a \in C(E^\top F)$, where a is a column vector of $\mathbb{R}^{\dim(S)}$. Similarly, for every out-domain instance, the feature projection $r_2 = E^\top Gb$ is used for some $b \in C(E^\top G)$, where b is a column vector of $\mathbb{R}^{\dim(U)}$. Therefore, the feature projections of the whole training dataset in the model space can be defined as $r = p_i r_1 + p_o r_2$. Assume $\mathcal{W}^*$ is the optimal set of model parameters, $\mathcal{W}^* = Er^*$, where $r^* = p_i E^\top Fa^* + p_o E^\top Gb*$, and a^*, b^* be the true feature projections.

Discussion (Data): In $S = C(X_{id}^\top)$ with basis F spanning X_{id}, $\forall\, x_i \in X_{id}, \exists\, z \in \mathbb{R}^{\dim(S)}$, $x = (Fz)^\top . X_{id} = (FZ)^\top$, where $Z = \{z\}$. Similarly, in $U = C(X_{ood}^\top)$ with basis G, $\forall\, x_{ood} \in X_{ood}$, $\exists\, v \in \mathbb{R}^{\dim(U)}$, $x_{ood} = (Gv)^\top . X_{ood} = (GV)^\top$, where $V = \{v\}$.

3.1.2 *Preliminary Optimization Target.*

Definition 3.2. Under the OOD setting, applying the same optimization objective to ID data with spurious features and out-domain data without the same spurious features is called undirected learning

Definition 3.3. Trained independently from scratch for the same number of iterations, the substructure within the original model having the best OOD generalization performance is defined as the OOD lottery [35].

For the structure-based approach searching the OOD lottery based on undirected learning, the optimization target can be formulated as:

$$\min \; \mathcal{L}(\mathcal{W}, X, Y) = \mathbb{E}_X \left\|X\mathcal{W} - Y\right\|_2^2 + \mathcal{S}\left(\mathcal{W}\right), \tag{1}$$

where $\mathcal{L}$ is the task-dependent loss function, and $\mathcal{S}$ is the function that induces the sparsity of the model structure to find the target subnetwork. The domain-generalized substructure is described by layer-wise channel saliencies in SFP. To this end, $\mathcal{S}$ is implemented by the squeeze-and-excitation module as suggested in [13]. The value of relevant parameters in t-th iteration is represented by subscript t, and the optimal value is represented by superscript $*$. Thus, the task loss in t-th iteration can be calculated as:

$$\mathcal{L}_t = \|X\mathcal{W}_t - Y\|_2^2 = \left\|X\mathcal{W}_t - X\mathcal{W}^*\right\|_2^2, \tag{2}$$

and the gradient is:

$$\frac{\partial \mathcal{L}_t}{\partial \mathcal{W}_t} = 2\left(\mathcal{W}_t - \mathcal{W}^*\right) X^\top X. \tag{3}$$

The orthogonal basis of the model space is regarded as the left singular vectors when performing SVD on the feature projections of data. The right singular vectors correspond to input data features, and the corresponding singular values can be defined as indicators of the importance of features of the current input w.r.t. the model structure. To internally observe the impact of ID and out-domain features on the model, the gradient accumulation is further transformed into a linear form:

$$\frac{\partial \mathcal{L}_t}{\partial \mathcal{W}_t} = 2(p_i^2(a_t - a^*)\Sigma^2_{E^\top F,t}X_{id} + p_o^2(b_t - b^*)\Sigma^2_{E^\top G,t}X_{ood}), \tag{4}$$

where Σ denotes the corresponding singular value matrix, and for simplicity, we omit t under Σ in the following discussion. The proof of Eq. Equation (4) is provided in [Appendix] A.1.

Since $\dim(U) = q \ll \dim(S) = p$, we have $\min \Sigma_{F^\top G} = \min \Sigma_{G^\top F} = \sigma^q_{G^\top F}$. Similarly, $\min \Sigma_{F^\top E} = \min \Sigma_{E^\top F} = \sigma^m_{E^\top F}$, and $\min \Sigma_{G^\top E} = \min \Sigma_{E^\top G} = \sigma^m_{E^\top G}$. Finally, the model parameters can be calculated as:

$$\begin{aligned}\mathcal{W}^{\infty} = \mathcal{W}_0 - 2lr \sum_{t=1}^{\infty}\sum_{i=1}^{m} p_i^2(a_t - a^*)\sigma^2_{E^\top F,t,i}X_{id} \\ - p_o^2(b_t - b^*)\sigma^2_{E^\top G,t,i}X_{ood}.\end{aligned} \tag{5}$$

3.1.3 Biased Performance on Out-domain and ID Data. Based on the gradient flow trajectories, we compare the learning process and final performance of the model for spurious and invariant features, respectively. We observe that the model structure obtained by undirected learning clearly differs in performance between ID data and out-of-domain data. With this observation, we propose the following propositions.

Proposition 3.4. *Undirected learning (full or sparse training) on biased data distributions can lead to significantly different forward speeds of the model learning along different data feature directions, and the difference has a second-order relationship with the proportion of different data distributions in the training set, i.e.:*

$$\left|\frac{\partial \mathcal{W}_t}{\partial(a_t - a^*)} - \frac{\partial \mathcal{W}_t}{\partial(b_t - b^*)}\right| \approx 2(p_i^2\Sigma^2_{E^\top F} - p_o^2\Sigma^2_{E^\top G}). \tag{6}$$

Discussion (Update Gradient): We compute the direction gradients along the directions of the feature projections of ID and out-domain data, respectively. As shown in Eq. 6, with $p_i \geq p_o$ in the context of OOD, the learning of the basis of the model space is gradually biased towards the directions of spurious features. By performing SVD on the projection of the basis vector of the feature space through the model space, the obtained singular value matrix can be regarded as the fitting degree of the model on the corresponding data distribution at t_{th} iteration.

Proposition 3.5. *Undirected learning (full or sparse training) on biased data distributions causes the model to be more biased towards training features with a larger proportion, bringing about significant performance differences in different data distributions, i.e.:*

$$\mathcal{L}_{ood} - \mathcal{L}_{id} \approx (p_i^2 - p_o^2)(1 - \Sigma_{F^\top G}) + \epsilon > 0, \tag{7}$$

where ϵ is the difference of initial feature projections between ID and out-domain data due to model initialization error. The full proof of Eq. 7 is provided in [Appendix] A.2.

Taking the risk difference between ID data and out-domain data of the trained model as the measurement of the OOD generalization, the following conclusion is derived, i.e.,

Corollary 3.6. *Undirected learning of networks on highly biased training domains (the dataset consists of a majority data group with spurious features) can only lead to substructures with sub-optimal OOD generalization performance.*

Discussion (Performance Difference): The result intuitively shows that the undirectly learned model performs better on feature distributions with larger instance numbers. As shown in Eq. 7, the difference in model performance between out-domain data and ID data is linearly related to the proportion of the corresponding instances and the correlation degree between the different feature distributions. Moreover, when the out-domain data has the same proportion as ID data in the training dataset (i.e., $p_i = p_o$) or the data distributions of them are consistent, the task loss difference between out-domain and ID data can be reduced to zero.

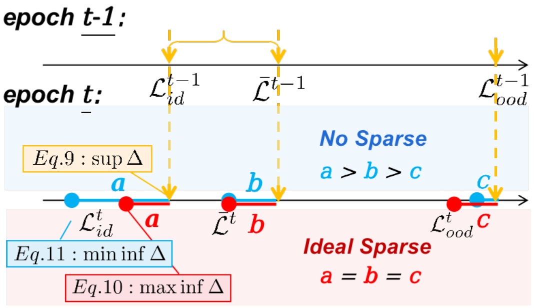


Figure 2: Identification of the ID instances dominated by spurious features. At epoch t, if no intervention is applied, the average loss drop on all data (blue b) should be smaller than that on ID data (blue a) and larger than that on out-domain data (blue c). The red line denotes an ideal regularization effect: the loss drops uniformly on all data.

3.2 SFP: An Spurious Feature-Targeted Model Pruning Method

To address the problem of sub-optimal OOD substructure caused by undirected training, we propose a novel method to effectively remove model branches that are only strongly correlated with spurious features. As demonstrated in Fig. 1, the pipeline consists of two stages, including spurious feature identification and model sparse training. Specifically, SFP identifies large spurious feature components within ID instances with high probability by observing the loss during training. It then can perform spurious feature-targeted model sparsity by analyzing the SVD of the feature projection matrix between the data and model space. We also provide a detailed theoretical analysis of both stages of the proposed SFP in the following part.

3.2.1 Spurious Feature Identification. As shown in Proposition. 3.5, if no intervention is applied, a model trained on a highly biased data distribution can be gradually biased towards ID data with lower prediction loss. Since the loss difference between ID and out-domain data can be approximately computed by $(p_i^2 - p_o^2)(1 - \sigma_{F^\top G})$, it is, therefore, can be adopted as the identification criterion for spurious features in each iteration. In brief, if the loss corresponding to the current data is lower than a threshold Δ, then the current data is likely to be an ID instance dominated by spurious features. Then we can further prune the spanning sets of model space along the directions of these spurious feature projections. To compute Δ, we first investigate the average loss in the $t-1$-th iteration as:

$$\bar{\mathcal{L}}^{t-1} \approx \mathcal{L}^{t-1}_{id} + p_o(p_i - p_o)(1 - \sigma^{t-1}_{F^\top G}). \tag{8}$$

As shown in Fig. 2, since $\mathcal{L}_{id}^{t} < \mathcal{L}_{id}^{t-1}$, we have:

$$\sup \mathcal{L}_{id}^{t} = |\bar{\mathcal{L}}^{t-1} - p_o(p_i - p_o)(1 - \sigma_{F^\top G}^{t-1})|. \tag{9}$$

Similar with Eq. 8, the lower bound of the loss on ID data at t-th iteration can be computed as:

$$\inf \mathcal{L}_{id}^{t} = |\bar{\mathcal{L}}^{t} - p_o(p_i - p_o)(1 - \sigma_{F^\top G}^{t})|. \tag{10}$$

The spurious feature-targeted regularization forces the model to learn invariant features and achieve fair loss reduction on all instances: $|\mathcal{L}_{id}^{t} - \mathcal{L}_{id}^{t-1}| = |\bar{\mathcal{L}}^{t} - \bar{\mathcal{L}}^{t-1}|$. Therefore, the ideal lower bound of the ID loss at t-th iteration is:

$$\begin{aligned} \inf \mathcal{L}_{id}^{t} = |&\bar{\mathcal{L}}^{t-1} - p_o(p_i - p_o)(1 - \sigma_{F^\top G}^{t-1})| \\ &- |\bar{\mathcal{L}}^{t} - \mathcal{L}^{\bar{t}-1}|. \end{aligned} \tag{11}$$

Thus, $\mathcal{L}_{id}^{t}$ is highly likely to be located in the range of $[\min \inf \mathcal{L}_{id}^{t}$, $\sup \mathcal{L}_{id}^{t}]$. The upper bound is used to compute Δ for identifying instances dominated by spurious features.

3.2.2 Spurious Feature-Targeted Pruning. SFP reacts to spurious feature-related instances by weakening their corresponding spurious feature projections into the model space, which can prevent the model from over-fitting on identified spurious features. To analyze the projections from data into the model space, we define $\Xi \in \mathbb{R}^{m \times m}$, $\Lambda \in \mathbb{R}^{p \times p}$, and $\Gamma \in \mathbb{R}^{q \times q}$ as the normalized orthogonal basis of $C(E^\top E)$, $C(E^\top F)$, and $C(E^\top G)$, spanning the optimal model projections, the feature projections of ID data into the model space, and the feature projections of out-domain data into the model space, respectively. ξ_i, λ_i, and γ_i denote the i-th column vectors in Ξ, Λ, and Γ, respectively. The following lemma illustrates the effectiveness of SFP, and its proof can be found in [Appendix] A.3.

Lemma 3.7. *Spurious feature-targeted model sparsity can effectively reduce the performance deviation of the learned model between in-domain data and out-domain data:*

$$\begin{aligned} &\left|\frac{\mathcal{R}(X_{ood}) - \mathcal{R}(X_{id})}{\mathcal{R}(X_{ood})^{sparse} - \mathcal{R}(X_{id})^{sparse}}\right| \\ \approx &\left|\frac{\sum_{j=1}^{m} p_o\tilde{\sigma}_j\xi_j\gamma_j X_{ood} - \sum_{i=1}^{m} p_i\sigma_i\xi_i\lambda_i X_{id}}{\sum_{j=1}^{m} p_o\tilde{\sigma}_j\xi_j\gamma_j X_{ood} - \sum_{i=1}^{\vartheta} p_i\sigma_i\xi_i\lambda_i X_{id}}\right| \geq 1, \end{aligned} \tag{12}$$

where $\mathcal{R}(\cdot)$ is the empirical risk function. σ_i and $\tilde{\sigma}_i$ is the i-th maximum in $\Sigma_{E^\top F}$ and $\Sigma_{E^\top G}$, and we have $\sigma > 0$ since the singular values are non-negative. m and ϑ are the rank of the singular value matrix after performing compact SVD and truncated SVD on the projections, respectively.

Proof of **Lemma 3.7**. As mentioned earlier, the projection space before the model sparsity can be represented as:

$$Er = \sum_{i=1}^{m} \left(p_i\sigma_i\xi_i\lambda_i^\top + p_o\tilde{\sigma}_i\xi_i\gamma_i^\top\right). \tag{13}$$

Specifically, SFP first performs SVD on the feature projections, which maps input data to a set of coordinates based on the orthogonal basis of model space. The matrices of left and right singular vectors correspond to the standard orthogonal basis of the model space and data space, respectively. The matrix of singular values corresponds to the direction weight of the action vectors in the projection matrix. SFP prunes the model by trimming the smallest singular values in Σ as well as their corresponding left and right singular vectors. In this way, SFP can remove the spurious features in ID data space and substructures in the model space simultaneously in a spurious feature-targeted manner along the directions with weaker actions for projection. Then, the projection space with only the most important ϑ singular values can be formalized as:

$$\begin{aligned} Er^{sparse} &= p_i\Xi\Sigma_{E^\top F}\Lambda^{-1} + p_o\xi\Sigma_{E^\top G}\Gamma^{-1} \\ &= \sum_{i=1}^{\vartheta} p_i\sigma_i\xi_i\lambda_i^\top + \sum_{j=1}^{m} p_o\tilde{\sigma}_j\xi_j\gamma_j^\top. \end{aligned} \tag{14}$$

Based on the representation of the projection spaces, the model response to data features $\mathcal{R}(X) = ErX$ can be calculated as:

$$\begin{aligned} \mathcal{R}(X) &= \left\{p_i\Xi\Sigma_{E^\top F}\Lambda^{-1} + p_o\xi\Sigma_{E^\top G}\Gamma^{-1}\right\}^\top X^\top \\ &= \sum_{i=1}^{m} \left\{p_i\sigma_i\xi_i\lambda_i^\top X^\top + p_o\tilde{\sigma}_i\xi_i\gamma_i^\top X^\top\right\}. \end{aligned} \tag{15}$$

□

3.3 Correspondence between Model Substructure and Spurious Features

In this section, we theoretically demonstrate that, with a reasonable setting of the sparse penalty for ID data, SFP can effectively reduce the overfitting of the model on spurious features while retaining the learning on invariant features. Specifically, we define $f^l(x)$ as the feature maps output of x at layer l. It represents the projection of x onto the model space defined over the spanning set E to be learned. We abbreviate the final probabilities as $f(x)$ for simplification. Referring to Sec. 3.2.1, we have $x \in X_{id}$ if $\mathcal{L}_{ce}(x) \leq \Delta$. Thus, the optimization target of SFP can be formulated as:

$$\min_{E} \mathbb{E}_{x \sim X}\mathcal{L}_{ce}(x, \mathcal{W}) + \eta \sum_{l=1}^{L} \mathbb{E}_{x \sim X_{id}}||f^l(x)||_1, \tag{16}$$

where η is the sparsity factor imposed on the feature projections for the identified ID data. Lemma 3.8 elucidates the setting of η. For a detailed proof, please refer to [Appendix] A.4.

Lemma 3.8. *Define $e = |f^*(x) - f(x)|$ as the l_1-norm between the groudtruth $f^*(x)$ and $f(x)$. When $\eta < 2e$, SFP can effectively reduce the learning of the model towards spurious features while keeping the performance on the other features.*

Proof of **Lemma 3.8:** The prediction errors of feature projections L_f can be defined as:

$$\begin{aligned} L_f &= |f^*(x) - f(x)|^2 \\ &= \sum_{i,j=j_1 \cup j_2} (f^*(x) - \sigma_{i,j_1}\xi_i \top \lambda_{j_1} - \sigma_{i,j_2}\xi_i \top \gamma_{j_2})^2, \end{aligned} \tag{17}$$

and the corresponding gradient is:

$$\begin{aligned} \frac{\partial L_f}{\partial \sigma_{i,j_1}\xi_i} &= \frac{\partial e^2}{\partial \sigma_{i,j}\xi_i} = 2e\frac{\partial e}{\partial \sigma_{i,j}\xi_i} \\ &= 2e\frac{\left|f^*(x) - \sigma_{i,j_1}\xi_i \top \lambda_{j_1} - \sigma_{i,j_2}\xi_i \top \gamma_{j_2})\right|}{\partial \sigma_{i,j}\xi_i} = -2e\lambda_{j_1}, \end{aligned} \tag{18}$$

where i and j are the index of column vectors in the orthogonal basis for model space and feature space, respectively. For out-domain data, the gradient of the column vectors in the OOD projection matrix interacting with the j_{th} feature vector is $-2e\gamma_{j_2}$. Then, split the in-domain features into spurious features F' and invariant features IN and out-domain features into unknown features G' and invariant features IN. With a high probability under the OOD setting, we assume F' and G' are orthogonal. To achieve the spurious feature-targeted unlearning and invariant feature-targeted learning of the model, we need to satisfy the following constraint:

$$2ep_i\lambda_{IN} + 2ep_o\gamma_{IN} - p_i\eta\lambda_{IN} > 2ep_o\gamma_{G'}$$
$$\Rightarrow \; \eta \le \frac{2ep_i\lambda_{IN} + 2ep_o\gamma_{IN} - 2ep_o\gamma_{G'}}{p_i\lambda_{IN}} \approx 2e. \tag{19}$$

□

Since the de-learning rate of the spurious feature is positively correlated with η, the upper bound $\eta = 2e$ is taken.

4 Experiments

In this section, we conducted extensive experiments on the DomainBed benchmarks [10] and other datasets that are widely used in the latest OOD studies. Due to space constraints, some experimental details are provided in [Appendix] B and C.

4.1 Experimental Setting

Datasets and Procedure. The proposed method is initially evaluated within the DomainBed framework using four datasets: ColoredMNIST (**CMNIST**), RotatedMNIST (**RMNIST**), as well as the multi-domain image classification datasets **PACS**, **OfficeHome**, **TerraInc**, and **DomainNet** [10]. To ensure comprehensive benchmarking, three synthetic datasets — FullColoredMNIST (**FCMNIST**), **ColoredObject**, and **SceneObject** — are included, along with two real-world image datasets, **CelebA** [22] and **WaterBirds** [31]. Fig. 3 illustrates the three synthetic datasets not encompassed within DomainBed, and more details are provided in [Appendix] C.1.

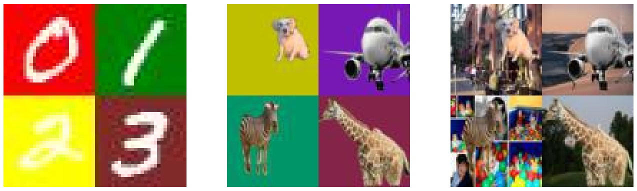

(a) FCMNIST (b) ColoredObject (c) SceneObject

Figure 3: Visualization of three synthetic OOD datasets.

Model and Implementation. To ensure a robust and equitable evaluation, the experimental settings in this work are consistent with the common practice established in antecedent studies. Specifically, for Rotated, Colored, and FCMNIST datasets, we use the 4-layer 3x3 ConvNet architecture as introduced in DomainBed. For the VLCS and PACS datasets, we utilize the ResNet-18 architecture as in IIB [18], with the default hyperparameters set in DomainBed. Additionally, for other larger datasets, we adopt the ResNet-50 architecture following the experimental settings outlined by previous works [27, 28]. All experiments are conducted on a workstation equipped with 8 Nvidia GTX 3090TI GPUs and a 3.6-GHZ Intel Core i9-9900KF CPU. The learning rate is initialized at 0.001 for digit datasets and 0.01 for object datasets. We employ the Adam optimizer for optimization in relatively simple image datasets, while SGD for more complex ones.

4.2 Comparison on DomainBed Benchmark

The experiment results on DomainBed demonstrate the superior performance of SFP over the state-of-the-art approaches. As shown in Table 1, SFP achieves the highest average accuracy of 72.8%, outperforming the benchmarked ERM (which is meticulously tuned within DomainBed and serves as a robust baseline) by 2.2%. On smaller datasets such as Colored and Rotated MNIST, most methods exhibit limited effectiveness. In contrast, SFP stands out by achieving an accuracy improvement of up to 14.0%, highlighting its robust feature-based recognition and suppression capabilities against correlation shifts. On larger datasets, SFP maintains satisfactory performance, demonstrating a remarkable accuracy increase up to 2.9% and 9.4% on VLCS and PACS, respectively. Notably, on the OfficeHome dataset, SFP boosts the OOD accuracy from 68.6% to 71.8%. The results also underscore the disadvantages of SOTAS in effectively addressing the correlation and diversity shifts simultaneously. For instance, while the ARM method excels in mitigating correlation shifts on Colored MNIST, it falters when confronted with diversity shifts in the OfficeHome dataset. Conversely, IIB performs well in scenarios involving diversity shifts but exhibits mediocre performance in correlation shift scenarios. Differently, SFP exhibits superior performance in most cases, emerging as a leading approach in the field of OOD generalization. More experimental details are provided in [Appendix] C.2.

4.3 Comparison on Other Benchmarks

We also conduct experiments on several widely-used datasets not included in DomainBed. For synthetic FCMNIST and ColoredObject datasets, bias coefficients (indicating the extent of data shift) are set as $(0.8, 0.6, 0.0)$. This implies that the digits in the two training domains are spuriously colored with probabilities of 0.8 and 0.6, while images in the test domain are randomly colored. For SceneObject dataset, we set the biased ratios as $(0.9, 0.7, 0.0)$, further hampering the model's capture of invariant features.

We compare SFP with the most comparable MRM, as well as their combined variants with IRM [2], V-REx [15], and DRO [28], on three synthetic datasets including FCMNIST, ColoredObject, and SceneObject. The results are shown in Table 2, demonstrating the superior performance of SFP under both independent and combined modes. To be specific, the results show that MRM compromises the generalization performance of the original algorithm in some cases. For example, the DRO algorithm independently achieves a test accuracy of 31.31% on SceneObject. However, when combined with MRM, the performance drops to 29.38%, while SFP contributes to an increased accuracy of 31.78%.

We also compare SFP with state-of-the-art SparseIRM [38] on FCMNIST with two different architectures, i.e., ResNet18 and MLP. Specifically, SFP outperforms SparseIRM with 3.41% higher test accuracy on MLP and even 29.12% on ResNet18. An interesting phenomenon is that, on small MLP, SparseIRM exhibits an obvious two-stage trend, which is consistent with regular non-feature-targeted

Table 1: DomainBed benchmark: Performance comparison (Accuracy %) between the proposed SFP method and the state-of-the-art domain generalization methods. "-" represents the missing data due to partially different settings. "Average" reports the average accuracy over all the datasets. We format first, second, and worse than ERM results.

Algorithm	CMNIST	RMNIST	VLCS	PACS	OfficeHome	TerraInc	DomainNet	Average
	MLP	MLP	ResNet-18	ResNet-18	ResNet-50	ResNet-50	ResNet-50	
ERM [30]	57.8±0.2	97.8±0.1	77.2±0.4	83.0±0.7	66.4±0.5	53.0±0.3	41.3±0.1	70.6
IRM [2]	67.7±1.2	97.5±0.2	76.3±0.6	81.5±0.8	63.0±2.7	50.5±0.7	28.0±5.1	69.0
GroupDRO [28]	61.1±0.9	97.9±0.1	77.9±0.5	83.5±0.2	66.2±0.6	52.4±0.1	33.4±0.3	67.5
Mixup [34]	58.4±0.2	98.0±0.1	77.7±0.6	83.2±0.4	68.0±0.2	54.4±0.3	39.6±0.1	63.3
MLDG [19]	58.2±0.4	97.8±0.1	77.2±0.9	82.9±1.7	66.6±0.3	52.0±0.1	41.6±0.1	68.0
MMD [1]	63.3±1.3	98.0±0.1	77.3±0.5	83.2±0.2	66.2±0.3	52.0±0.4	23.5±9.4	66.2
CDANN [21]	59.5±2.0	97.9±0.0	77.5±0.2	78.8±2.2	65.3±0.5	50.8±0.6	38.5±0.2	66.9
MTL [3]	57.6±0.3	97.9±0.1	76.6±0.5	83.7±0.4	66.5±0.4	52.2±0.4	40.8±0.1	67.9
SagNet [24]	58.2±0.3	97.9±0.0	77.5±0.3	82.3±0.1	67.5±0.2	52.5±0.4	40.8±0.2	68.1
ARM [36]	63.2±0.7	98.1±0.1	76.6±0.5	81.7±0.2	64.8±0.4	51.2±0.5	36.0±0.2	67.4
V-REx [15]	67.0±1.3	97.9±0.1	76.7±1.0	81.3±0.9	65.7±0.3	51.4±0.5	30.1±3.7	67.2
RSC [14]	58.5±0.5	97.6±0.1	77.5±0.5	82.6±0.7	66.5±0.6	52.1±0.2	38.9±0.6	67.7
Fishr [27]	68.8±1.4	97.8±0.1	-	-	68.2±0.2	53.6±0.4	**41.8**±0.2	-
SFP	**71.6**±0.3	**98.3**±1.4	**79.2**±0.7	**90.7**±0.1	**71.8**±0.1	57.8±0.3	40.0±0.7	**72.8**

Table 2: OOD generalization performance on FullColoredMNIST, ColoredObject, and SceneObject. "MRM+X" and "SFP+X" indicate the integration of MRM/SFP in the "X" algorithm. The "Unbiased" row reports the original accuracy for each dataset without data distribution shifts.

Method	FCMNIST	ColoredObject	SceneObject
ERM	62.2	59.2	27.4
MRM	81.0	60.7	26.7
SFP	**84.3**	**61.01**	**28.4**
IRM	78.0	62.9	36.9
MRM +IRM	89.3	64.5	36.9
SFP+IRM	**89.9**	**65.8**	**38.1**
V-REx	87.8	64.7	36.7
MRM +V-REx	92.2	64.5	36.7
SFP+V-REx	93.4	66.1	37.9
DRO	62.9	66.8	31.3
MRM +DRO	80.5	66.2	29.4
SFP+DRO	**85.2**	**68.4**	**31.8**
UNBIASED	94.0	75.8	45.5

model pruning. Differently, SFP consistently shows a stable learning process and achieves higher performance in both ID (train) and OOD (test) environments. Due to space constraints, the experimental details are provided in [Appendix] C.3.

4.4 Ablation Study

Loss Tracking. We visualize and compare loss values between ERM and our proposed SFP to assess the efficacy of our introduced regularization term. As shown in Fig. 4, throughout the training process, the loss of in-domain (ID) instances consistently remains lower than that of out-domain instances, validating Proposition 3.5. In ERM, the rapid convergence of ID instance loss (depicted by red lines) indicates an excessive focus on biased data, leading to overfitting spurious features and neglecting invariant features. Conversely, in SFP, the gap between loss values for ID and out-domain instances narrows significantly, underscoring the effectiveness of spurious feature-targeted pruning. What's more, the optimization of SFP won't hinder convergence speed as well as adversely affects the performance of ID instances.

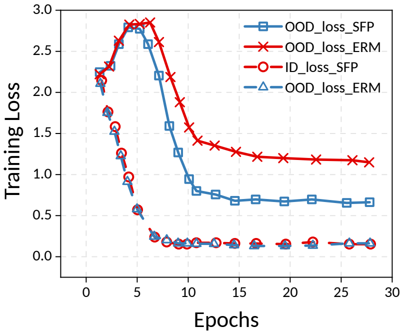
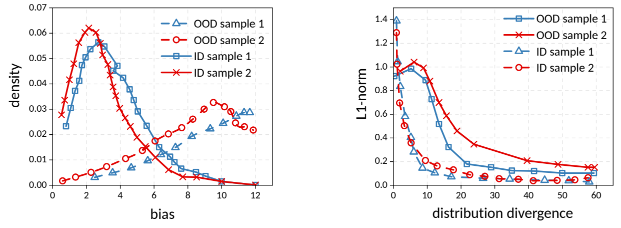


Figure 4: Training loss visualization

Prediction Confidence. The inherent motivation of SFP originates from scrutinizing the behavioral disparities between ID samples and OOD samples under ERM, which is illustrated via two empirical experiments as follows. We first measure the bias between the maximum value and other values in the logits vectors corresponding to different samples, where the maximum typically represents the prediction. A large logits discrepancy suggests a



(a) Bias density

(b) Distribution divergence

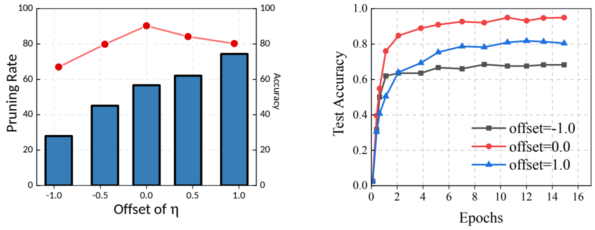
Figure 5: The probability density of bias between the max value and the others in predicted distribution.

significant divergence between the probability densities of the predicted class and others, which can be used as a metric for gauging the prediction confidence. The results, depicted in Fig. 5a, reveal that ID samples generally exhibit larger logits discrepancies compared to OOD samples, indicating a tendency of the current model to allocate greater confidence to the predictions of ID samples.

Additionally, we evaluate the l_1-norm between the predicted and true distributions over different classes to gauge the degree of the model capturing different features. The results are shown in Fig. 5b. It's evident that the distribution loss of ID samples sharply decreases in the early training stages but gradually slows down afterward. Conversely, OOD samples initially show a slight increase in distribution loss, followed by a steep decrease. This early training behavior suggests that the model initially prioritizes spurious correlations, but as training progresses, SFP mitigates the fit of spurious correlations while promoting the learning of invariant features. As a result, the downward trend of distribution loss for ID samples decelerates, while the trend for OOD samples starts to rise.

Sparsity Analysis. Prior structure-based OOD studies usually utilize human-crafted hyperparameters to find a suitable functional OOD substructure. In contrast, our method treats the sparsity coefficients (Δ, η) as dynamic variables that are calculated dynamically during training, i.e., *the proposed SFP intelligently determines the optimal OOD sparsity and structure based on inherent data attributes.* Specifically, (Δ) gives a sparsity threshold based on inherent statistical and geometric biases within the data (e.g., Eqs. 9-11), and η adjusts the penalty strength based on dynamic training feedback (e.g., Eq. 6). To empirically evaluate the sparsity of our model and, at the same time, provide a quantitative impact of η on OOD accuracy, we conduct experiments on varied offsets to the theoretically computed η (2e). Specifically, the offsets are ranged in [-1.0, -0.5, 0.0, 0.5, 1.0]. The results regarding model sparsity and test accuracy are shown in Fig. 6. The corresponding OOD accuracy are [73.01262%, 79.84853%, 86.30715%, 84.19074%, 76.23703%], and the pruning rates are ranged in [27.94951%, 45.09116%, 56.70407%, 62.09122%, 74.40112%]. The results demonstrate that the autonomous acquisition of sparsity and sparse structures (offset of 0) yields superior OOD performance than empirical sparse settings.

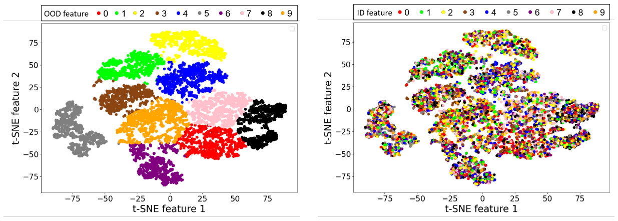
Feature Visualization. We visualize the extracted features using t-distributed stochastic neighbor embedding (t-SNE) for dimensionality reduction to explore the SFP model's learned representations. Experiments are conducted on FullColoredMNIST datasets. The models are trained on domain-related samples and tested on domain-unrelated samples with random colors. The results are shown in Fig. 7, where each data point represents an image. Notably, the spatial arrangement corresponds to the reduced shape features. The features cluster into ten groups. The left subplots color each point based on invariant features, i.e., samples with the same digit are colored identically. For example, as shown in Fig. 7a, each cluster contains points belonging to one class. Conversely, all right subplots color each point based on spurious features, where samples with the same spurious feature (e.g., red 2 and red 3) are colored identically. The results are shown in Fig. 7b, each cluster (class) involves diverse spurious features, indicating that the current classification results are independent of spurious features. This suggests that clustered features are specific to invariant digit shapes and remain unaffected by color variations, demonstrating that SFP could successfully acquire disentangled representations.



(a) Model sparsity

(b) Test accuracy

Figure 6: The effect of different η values on the model sparsity and accuracy.



(a)

(b)

Figure 7: The visualization of the features learned by SFP.

5 Conclusion

In this paper, we introduce a novel spurious feature-targeted model pruning framework, dubbed SFP, designed to automatically explore the optimal model substructure for improved out-of-distribution (OOD) generalization. By effectively identifying spurious features within in-distribution (ID) instances during training, SFP can selectively remove model branches that heavily depend on these spurious features. As a result, SFP attenuates the impact of spurious features on the model's representation space and guides the model learning process toward invariant features. Additionally, we provide a detailed theoretical analysis to establish the rationality of our approach and offer a proof framework for understanding OOD structures via model sparsity. Experimental results corroborate the effectiveness of our proposed method.

6 Acknowledgment

This research was supported by funding from the Hong Kong RGC General Research Fund (No. 152211/23E, 152164/24E, 152203/20E, 152244/21E, 152169/22E, 152228/23E), Research Impact Fund (No. R5060-19, No. R5034-18), Areas of Excellence Scheme (AoE/E-601/22-R), the National Natural Science Foundation of China (No. 62102327, 62192781, 62172326, and 62137002), the PolyU Internal Fund (No. P0043932), the Key-Area Research and Development Program of Guangdong Province (No. 2021B0101400003), and the Project of China Knowledge Centre for Engineering Science and Technology. This research was also supported by NVIDIA AI Technology Center.

References

[1] Kei Akuzawa, Yusuke Iwasawa, and Yutaka Matsuo. 2019. Adversarial Invariant Feature Learning with Accuracy Constraint for Domain Generalization. In *Machine Learning and Knowledge Discovery in Databases - European Conference, ECML PKDD 2019, Würzburg, Germany, September 16-20, 2019, Proceedings, Part II (Lecture Notes in Computer Science, Vol. 11907)*. Springer, 315–331.

[2] Martin Arjovsky, Léon Bottou, Ishaan Gulrajani, and David Lopez-Paz. 2019. Invariant risk minimization. *arXiv preprint arXiv:1907.02893* (2019).

[3] Gilles Blanchard, Aniket Anand Deshmukh, Ürün Dogan, Gyemin Lee, and Clayton D. Scott. 2017. Domain Generalization by Marginal Transfer Learning. *J. Mach. Learn. Res.* 22 (2017), 2:1–2:55. https://api.semanticscholar.org/CorpusID:59362358

[4] Yongqiang Chen, Wei Huang, Kaiwen Zhou, Yatao Bian, Bo Han, and James Cheng. 2023. Understanding and Improving Feature Learning for Out-of-Distribution Generalization. In *Thirty-seventh Conference on Neural Information Processing Systems*.

[5] Yongqiang Chen, Kaiwen Zhou, Yatao Bian, Binghui Xie, Bingzhe Wu, Yonggang Zhang, Kaili Ma, Han Yang, Peilin Zhao, Bo Han, and James Cheng. 2023. Pareto Invariant Risk Minimization: Towards Mitigating the Optimization Dilemma in Out-of-Distribution Generalization. In *The Eleventh International Conference on Learning Representations, ICLR 2023, Kigali, Rwanda, May 1-5, 2023*. OpenReview.net.

[6] Matthew Cook, Alina Zare, and Paul Gader. 2020. Outlier detection through null space analysis of neural networks. *arXiv preprint arXiv:2007.01263* (2020).

[7] Simon S Du, Wei Hu, and Jason D Lee. 2018. Algorithmic regularization in learning deep homogeneous models: Layers are automatically balanced. *Advances in neural information processing systems* 31 (2018).

[8] Yanrui Du, Jing Yan, Yan Chen, Jing Liu, Sendong Zhao, Qiaoqiao She, Hua Wu, Haifeng Wang, and Bing Qin. 2023. Less Learn Shortcut: Analyzing and Mitigating Learning of Spurious Feature-Label Correlation. In *Proceedings of the Thirty-Second International Joint Conference on Artificial Intelligence, IJCAI 2023, 19th-25th August 2023, Macao, SAR, China*. ijcai.org, 5039–5048.

[9] Shaohua Fan, Xiao Wang, Chuan Shi, Peng Cui, and Bai Wang. 2024. Generalizing Graph Neural Networks on Out-of-Distribution Graphs. *IEEE Trans. Pattern Anal. Mach. Intell.* 46, 1 (2024), 322–337.

[10] Ishaan Gulrajani and David Lopez-Paz. 2021. In Search of Lost Domain Generalization. In *International Conference on Learning Representations*. https://openreview.net/forum?id=lQdXeXDoWtI

[11] Song Han, Jeff Pool, John Tran, and William J. Dally. 2015. Learning both Weights and Connections for Efficient Neural Network. In *NIPS*. 1135–1143.

[12] Yang He, Guoliang Kang, Xuanyi Dong, Yanwei Fu, and Yi Yang. 2018. Soft Filter Pruning for Accelerating Deep Convolutional Neural Networks. In *IJCAI*. ijcai.org, 2234–2240.

[13] Jie Hu, Li Shen, and Gang Sun. 2017. Squeeze-and-Excitation Networks. *CoRR* abs/1709.01507 (2017).

[14] Zeyi Huang, Haohan Wang, Eric P. Xing, and Dong Huang. 2020. Self-Challenging Improves Cross-Domain Generalization. *ArXiv* abs/2007.02454 (2020). https://api.semanticscholar.org/CorpusID:220363892

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

[16] Ananya Kumar, Aditi Raghunathan, Robbie Jones, Tengyu Ma, and Percy Liang. 2022. Fine-Tuning can Distort Pretrained Features and Underperform Out-of-Distribution. arXiv:2202.10054 [cs.LG]

[17] Yann LeCun, John S. Denker, and Sara A. Solla. 1989. Optimal Brain Damage. In *NIPS*. Morgan Kaufmann, 598–605.

[18] Bo Li, Yifei Shen, Yezhen Wang, Wenzhen Zhu, Dongsheng Li, Kurt Keutzer, and Han Zhao. 2022. Invariant information bottleneck for domain generalization. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 36. 7399–7407.

[19] Da Li, Yongxin Yang, Yi-Zhe Song, and Timothy M. Hospedales. 2018. Learning to Generalize: Meta-Learning for Domain Generalization. In *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence, (AAAI-18), the 30th innovative Applications of Artificial Intelligence (IAAI-18), and the 8th AAAI Symposium on Educational Advances in Artificial Intelligence (EAAI-18), New Orleans, Louisiana, USA, February 2-7, 2018*. AAAI Press, 3490–3497.

[20] Hao Li, Asim Kadav, Igor Durdanovic, Hanan Samet, and Hans Peter Graf. 2017. Pruning Filters for Efficient ConvNets. In *ICLR (Poster)*. OpenReview.net.

[21] Ya Li, Xinmei Tian, Mingming Gong, Yajing Liu, Tongliang Liu, Kun Zhang, and Dacheng Tao. 2018. Deep Domain Generalization via Conditional Invariant Adversarial Networks. In *European Conference on Computer Vision*. https://api.semanticscholar.org/CorpusID:52956008

[22] Ziwei Liu, Ping Luo, Xiaogang Wang, and Xiaoou Tang. 2015. Deep Learning Face Attributes in the Wild. In *Proceedings of International Conference on Computer Vision (ICCV)*.

[23] Vaishnavh Nagarajan, Anders Andreassen, and Behnam Neyshabur. 2021. Understanding the failure modes of out-of-distribution generalization. In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*. OpenReview.net.

[24] Hyeonseob Nam, Hyunjae Lee, Jongchan Park, Wonjun Yoon, and Donggeun Yoo. 2019. Reducing Domain Gap via Style-Agnostic Networks. *ArXiv* abs/1910.11645 (2019). https://api.semanticscholar.org/CorpusID:204803849

[25] Behnam Neyshabur, Ryota Tomioka, and Nathan Srebro. 2014. In search of the real inductive bias: On the role of implicit regularization in deep learning. *arXiv preprint arXiv:1412.6614* (2014).

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

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

[28] Shiori Sagawa, Pang Wei Koh, Tatsunori B. Hashimoto, and Percy Liang. 2019. Distributionally Robust Neural Networks for Group Shifts: On the Importance of Regularization for Worst-Case Generalization. *CoRR* abs/1911.08731 (2019).

[29] Shiori Sagawa, Aditi Raghunathan, Pang Wei Koh, and Percy Liang. 2020. An investigation of why overparameterization exacerbates spurious correlations. In *International Conference on Machine Learning*. PMLR, 8346–8356.

[30] V.N. Vapnik. 1999. An overview of statistical learning theory. *IEEE Transactions on Neural Networks* 10, 5 (1999), 988–999. https://doi.org/10.1109/72.788640

[31] C. Wah, S. Branson, P. Welinder, P. Perona, and S. Belongie. 2011. *The Caltech-UCSD Birds-200-2011 datase*. Technical Report CNS-TR-2011-001. California Institute of Technology.

[32] Haoqi Wang, Zhizhong Li, Litong Feng, and Wayne Zhang. 2022. Vim: Out-of-distribution with virtual-logit matching. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 4921–4930.

[33] Wei Wen, Chunpeng Wu, Yandan Wang, Yiran Chen, and Hai Li. 2016. Learning Structured Sparsity in Deep Neural Networks. In *NIPS*. 2074–2082.

[34] Shen Yan, Huan Song, Nanxiang Li, Lincan Zou, and Liu Ren. 2020. Improve Unsupervised Domain Adaptation with Mixup Training. *CoRR* abs/2001.00677 (2020).

[35] Dinghuai Zhang, Kartik Ahuja, Yilun Xu, Yisen Wang, and Aaron Courville. 2021. Can subnetwork structure be the key to out-of-distribution generalization?. In *International Conference on Machine Learning*. PMLR, 12356–12367.

[36] Marvin Zhang, Henrik Marklund, Abhishek Gupta, Sergey Levine, and Chelsea Finn. 2020. Adaptive Risk Minimization: A Meta-Learning Approach for Tackling Group Shift. *CoRR* abs/2007.02931 (2020).

[37] Chenglong Zhao, Bingbing Ni, Jian Zhang, Qiwei Zhao, Wenjun Zhang, and Qi Tian. 2019. Variational Convolutional Neural Network Pruning. In *CVPR*. Computer Vision Foundation / IEEE, 2780–2789.

[38] Xiao Zhou, Yong Lin, Weizhong Zhang, and Tong Zhang. 2022. Sparse Invariant Risk Minimization. In *Proceedings of the 39th International Conference on Machine Learning (Proceedings of Machine Learning Research, Vol. 162)*, Kamalika Chaudhuri, Stefanie Jegelka, Le Song, Csaba Szepesvari, Gang Niu, and Sivan Sabato (Eds.). PMLR, 27222–27244.