Rich Feature Learning via Diversification

Xi Leng[1] Yongqiang Chen[2] Xiaoying Tang[1,3,4*] Yatao Bian[5]
[1]School of Science and Engineering, The Chinese University of Hong Kong, Shenzhen, Guangdong 518172, China
[2]The Chinese University of Hong Kong, Hong Kong SAR, China
[3]Shenzhen Future Network of Intelligence Institute (FNii-Shenzhen)
[4]Guangdong Provincial Key Laboratory of Future Networks of Intelligence
[5]Department of Computer Science, National University of Singapore, Singapore
xileng@link.cuhk.edu.cn, yqchen@cse.cuhk.edu.hk, tangxiaoying@cuhk.edu.cn, ybian@nus.edu.sg

Abstract

Rich Feature Learning (RFL) aims to extract all beneficial features from the training distribution and shows promise for Out-of-Distribution (OOD) generalization. Despite its success, a precise and comprehensive definition of "richness" remains elusive. Through an in-depth comparison between RFL and empirical risk minimization (ERM), we identify that feature diversity is the key differentiator driving RFL's superior OOD performance. Building on this insight, we contribute a formal definition of rich features, encompassing both informativeness and diversity. Leveraging this foundation, we propose ***D****iversity-f****O****unded* ***R****ich f****E****ature l****E****arni****N****g (*DOREEN*), a simple yet highly effective RFL algorithm that trains multiple models with identical architectures concurrently to promote feature diversity. We theoretically demonstrate that* DOREEN *not only realizes the benefits of RFL but also addresses the limitations of prior RFL algorithms. Extensive experiments validate that* DOREEN *learns richer features and consistently enhances OOD performance across various OOD objectives.*

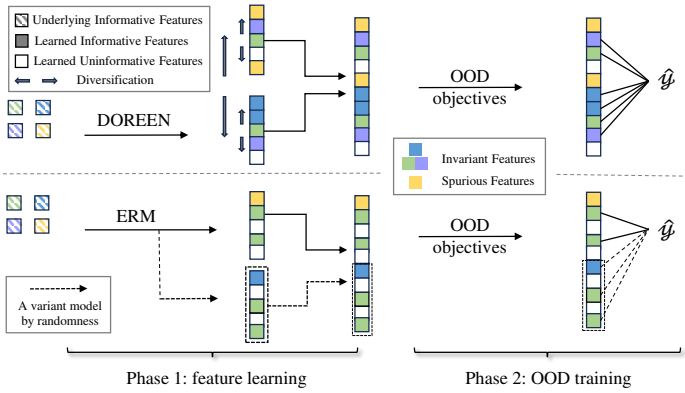
1. Introduction

The significant performance degradation of models trained with ERM on OOD data is commonly attributed to learning spurious features [4, 6]. To address this challenge, there has been a surge of efforts in developing OOD objectives to regularize ERM feature learning [2, 4, 11, 26, 27, 37, 44]. However, such regularization can disrupt the standard ERM feature learning process, introducing a substantial optimization dilemma [10, 52]. To overcome these optimization difficulties, the concept of Rich Feature Learning (RFL) was introduced [10, 52]. As illustrated in Figure 1(a), RFL focuses on training a rich featurizer Φ during Phase 1, which extracts a broader and more comprehensive set of features from the training data compared to ERM. This featurizer lays the groundwork for Phase 2, where a simple (often linear) classifier ω is trained on Φ to yield the final model $\omega \cdot \Phi$. Despite its success, RFL still lacks a formal and clear definition of "rich features".

Our experiments, approached through the lens of diversity, delve into what distinguishes "rich features" from those learned via ERM. Specifically, we assess the diversity of the features learned by ERM together with two RFL algorithms: BONSAI (also called RFC) [52] and FeAT [10] using Vendi Score [15] and examine the corresponding OOD performance. In Sec. 3.1 we compare the diversity and OOD performance of the features learned by ERM with those of BONSAI and FeAT. The results indicate that ERM-trained featurizers exhibit lower feature diversity and inferior OOD performance compared to RFL methods. In Sec. 3.2 we provide a comprehensive analysis of ERM's training process, observing that the OOD performance closely aligns with the feature diversity, which initially rises briefly before consistently declining. We also find that, while the number of learned features increases early in the training, this growth quickly plateaus and the intrinsic similarity within each feature intensifies over time.

These empirical findings collectively underscore that feature diversity is the primary factor distinguishing RFL from ERM. Building on this insight, we propose a formal definition of rich features as those that are both *diverse* and *informative* (Sec. 3.3). We then show theoretically that the existing SOTA RFL methods can fail catastrophically under strong spurious correlations (Sec. 4.1). Consequently, we propose DOREEN (**D**iversity-f**O**unded **R**ich f**E**ature l**E**arni**N**g), a simple yet powerful RFL framework (Sec. 4.2). DOREEN trains multiple models with identical

*Corresponding author.



(a)

Algorithm	RFL	Diversity Usage	Implementation	Scope
ERM	✗	✗	N/A	N/A
BONSAI [52]	✓	✗	N/A	N/A
FeAT [10]	✓	✗	N/A	N/A
DiWA [38]	✗	✓	implicit	whole model
Deep Emsemble [30]	✗	✓	implicit	whole model
Teney et al. [46]	✗	✓	explicit	classifier
DivDis [31]	✗	✓	explicit	classifier
D-BAT [34]	✗	✓	explicit	whole model
WLD-Reg [29]	✗	✓	explicit	single featurizer
DOREEN (Ours)	✓	✓	explicit	multiple featurizer

(b)

Figure 1. (a) Illustration of DOREEN vs. ERM: DOREEN leverages intra-model and inter-model diversity losses to encourage the learning of richer features compared to ERM, providing a stronger foundation for OOD training. (b) Overview of diversity-based algorithms for OOD generalization: Algorithms are classified as "explicitly" if they include a diversity penalty.

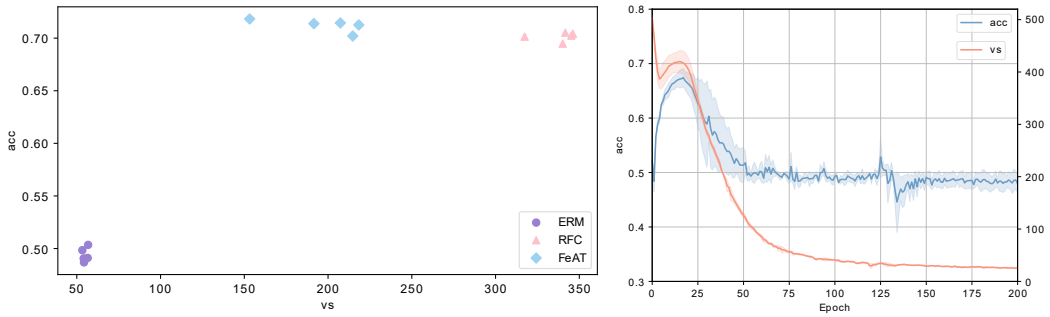
structures, each optimized with individual ERM losses and a shared diversity loss that promotes both intra-model and inter-model diversity. The final rich feature representation is obtained by concatenating the featurizers of the trained models, as illustrated in Figure 1 (a). Crucially, DOREEN is agnostic to the specific diversity metric used—serving as a flexible bridge between general diversity principles and rich feature learning. This design allows practitioners to plug in the diversity measure best suited to their data, domain, or constraints. By embedding diversity directly into the learning process, DOREEN provably incorporates richer features than ERM (Sec. 4.3) and effectively addresses scenarios where existing RFL methods struggle. Extensive experiments (Sec. 5) demonstrate that DOREEN not only significantly outperforms ERM but also matches or surpasses existing RFL algorithms.

We summarize our contributions as follows: 1) Key observation: We identify feature diversity as a critical factor for cultivating rich features. 2) A general RFL framework: We propose a formal definition for "rich features" and introduce DOREEN —a framework that explicitly promotes intra- and inter-model diversity while remaining agnostic to the choice of diversity metric. We provide theoretical evidence for its ability to incorporate richer features than ERM and address fundamental limitations of existing RFL methods. 3) Empirical validation: We validate DOREEN across various settings, demonstrating that it significantly improves upon ERM and rivals or surpasses existing RFL algorithms. Moreover, we verify that DOREEN effectively handles scenarios where existing RFL methods may encounter challenges.

2. Related Work

OOD Generalization. Empirical Risk Minimization (ERM) has long been criticized for its failure in Out-of-Distribution (OOD) generalization due to its reliance on spurious features. This has spurred extensive research to develop OOD objectives that foster invariant feature learning robust to distribution shifts. These methods leverage multiple training sets (domains) to simulate potential distributional changes and enforce feature invariance across them [2, 4, 9, 26, 27, 37, 44]. However, the optimization challenges posed by these objectives often surpass the complexities of ERM [11] and introduce a severe disturbance to the ERM feature learning [52], leading to empirical observations by Zhang et al. [52] that question the effectiveness of these OOD objectives in real-world tasks. The difficulty lies in striking the fine balance: overly restrictive objectives necessitate ERM pre-training and precise hyperparameter adjustments, while overly permissive ones fail to preserve invariant features, potentially causing model degeneration. As a result, ERM shows superior performance over most OOD objectives [39] in standard benchmarks such as DomainBed [16], DrugOOD [21] and WILDS [25]. In contrast to these approaches, the concept of Rich Feature Learning (RFL) has emerged. It aims to develop representations that encapsulate a broader spectrum of useful features, offering novel insights into enhancing OOD performance.

Rich Feature Learning (RFL). RFL paradigm separates the training process for OOD generalization tasks into two complementary phases: feature learning and OOD training, as illustrated in Figure 1 (a). Within this process, RFL focuses on developing comprehensive feature representations including both invariant and spurious features. Then in Phase 2 the OOD objectives are applied to the learned features to effectively identify and utilize the invariant components for prediction. This two-phase structure enables effective OOD generalization by decoupling feature learning from invariance enforcement. BONSAI [52] constructs rich representations by iteratively learning new features from incorrectly predicted subsets (augmentation sets) while retaining previously learned features that



(a) (b)

Figure 2. The empirical results on ColoredMNIST-025. Post featurizer training, we measure the feature diversity using the Vendi Score (vs), freeze the featurizer, and subsequently train a classifier using V-REX for OOD performance (acc) assessment. (a): The feature diversity (x-axis) and OOD performance (y-axis) of featurizers trained with ERM and two RFL algorithms over five different random seeds. (b): ERM training dynamics over three different random seeds. The x-axis illustrates the evaluation epochs of the featurizer. The y-axis displays the corresponding OOD accuracy and Vendi Score values at each evaluation epoch.

correctly predict data segments (retention sets). While effective, BONSAI relies on multiple initializations of the whole network and the final intricate process to integrate insights from all models demands numerous training epochs to achieve convergence. Alternatively, FeAT [10] efficiently learns rich representations by optimizing a combined loss that includes an ERM loss on the retention sets and a DRO loss [33] on the augmentation sets. Both BONSAI and FeAT exhibit superior OOD performance compared to ERM, suggesting their success in fostering richer representations. However, they lack a precise definition of "richness", obtaining a broader spectrum of useful features by training over multiple rounds on reweighted datasets. Moreover, their effectiveness is contingent on the quality of the augmentation set. For instance, if the training set is dominated by a strong, easily detectable correlation, the augmentation set becomes negligible, limiting BONSAI and FeAT's capacity to discern additional informative features. The situation deteriorates if this dominant correlation is spurious, such as the background in the waterbirds dataset [41], where the objective is to distinguish bird breeds, or drill artifacts in the chest-x-ray dataset [13], where the task is to identify diseases. These limitations are theoretically analyzed in Proposition 1 and empirically validated in Tab. 1.

Leveraging Diversity for Enhanced OOD Generalization. Allen-Zhu and Li [3] argue that many parameter configurations can explain finite training data equally well—referred to as "multi-view"—and emphasize the importance of capturing this diversity for robust OOD generalization [22]. Multi-view structures are prevalent in real-world scenarios; for example, lions can be recognized either by their manes or facial features. Capturing such diverse cues enables models to generalize robustly, even to female lions that lack manes. There has been a surge of efforts aimed at leveraging diversity to enhance OOD generalization. Rame et al. [38] propose weight averaging across multiple models to encourage diversity. However, this approach does not explicitly promote diversity and relies on the randomness of initialization to generate model variance, which may lead to redundancy [36]. Lee et al. [31], Pagliardini et al. [34], Teney et al. [46] opt to promote diversity through disagreement at the output level (such as output distribution or gradient). DOREEN stands out by explicitly enforcing diversity at the feature level, enabling the model to learn a broader spectrum of features that generalize better to unseen distributions. While Laakom et al. [29] introduces diversity within a single feature extractor, our method extends this by incorporating inter-model diversity—an enhancement both theoretically grounded and empirically validated to yield richer representations and stronger OOD performance. Moreover, Benoit et al. [8] reveals that diversity alone is insufficient for OOD generalization, as it may lead to reliance on spurious features. Built on the RFL framework, DOREEN can leverage the OOD training phase to select invariant features, effectively addressing this limitation.

Comparison to previous works. A detailed comparison of DOREEN with the most relevant algorithms is presented in Figure 1 (b). Owing to space limitations, extended discussion of related work is deferred to Sec. 8. As an RFL method, DOREEN explicitly promotes diversity within the feature space at both inter-model and intra-model levels, enabling the learning of richer representations to improve OOD generalization.

3. Motivating Studies and Featurizer Richness

To explore the distinctive aspects of "rich features" compared to those learned by ERM, we initiated a series of experiments focusing on feature diversity utilizing the ColoredMNIST dataset [4] (denoted as ColoredMNIST-025). We also included a modified version named ColoredMNIST-01. The primary distinction between these two datasets lies in the feature-label correlation: spurious (ColoredMNIST-025) or invariant (ColoredMNIST-01) features are better correlated with labels. Due to

the space limit, the results on ColoredMNIST-01 are shown in Figure 5. Detailed information about the ColoredMNIST dataset and empirical configurations is provided in Sec. 10.1.

3.1. Comparing ERM and RFL Algorithms

We begin our analysis by comparing the feature diversity and OOD performance between ERM and two SOTA RFL algorithms: BONSAI (also known as RFC) [52] and FeAT [10]. To assess feature diversity, we utilize the Vendi Score, as conceptualized by Friedman and Dieng [15].

Let $x_1, x_2, ..., x_n \in \mathfrak{X}$ denote a collection of samples, $k : \mathfrak{X} \times \mathfrak{X} \mapsto \mathbb{R}$ be a positive semidefinite similarity function, and $k(x, x) = 1$ for all x, $\mathbf{K} \in \mathbb{R}^{n \times n}$ is a kernel matrix with entry $\mathbf{K}_{i,j} = k(x_i, x_j)$. Then the Vendi Score (VS) is defined as $VS_k(x_1, \ldots, x_n) = \exp\left(-\sum_{i=1}^{n} \lambda_i \log \lambda_i\right)$ where $\lambda_1, \lambda_2, ..., \lambda_n$ are the eigenvalues of $\mathbf{K}/n$. In practice, after training the featurizer, we gather its outputs using a small, randomly selected subset of the training data. These outputs are then employed to construct a similarity matrix that measures the mutual similarity between every pair of dimensions, from which we calculate the Vendi Score. Subsequently, we freeze the featurizer and utilize V-REX [27] to train a classifier on top of it to evaluate the OOD performance, as V-REX is a SOTA OOD objective that can better showcase the quality of the learned features. The results in Figure 2 (a) reveal a marked difference in feature diversity.

Features learned through ERM exhibit significantly lower diversity compared to those obtained via RFC and FeAT. This, in turn, leads to a notably lower OOD accuracy for ERM when these features are applied for inference in contrast to the performance achieved by RFC and FeAT.

3.2. Study of ERM Training Dynamics

We also conducted experiments to track the evolution of the feature diversity and OOD performance throughout the training process of a featurizer trained with ERM. Over the course of 1,000 epochs, the featurizer trained with ERM was evaluated every 5 epochs. At these intervals, we recorded the Vendi Score of the featurizer, then froze it to train a new classifier using V-REX, subsequently measuring the OOD accuracy. The results of these experiments are detailed in Figure 2 (b). Initially, due to the random initialization of the featurizer, a wide variety of random features are created, resulting in an almost-maximal Vendi Score. During the early stages of training, there is a pronounced synchrony between the rise in feature diversity and the improvement in OOD accuracy, indicating that the featurizer is rapidly learning diverse and informative features. ERM-trained featurizers transiently possess high feature diversity and show promising OOD performance. Yet, this diversity diminishes as training advances, resulting in a parallel decrease in OOD accuracy. We also discern the features learned by ERM during the training period, the results in Figure 6 further validate the point.

3.3. Feature Richness Formulation

The above experiments highlight the essence of RFL: acquiring diverse representations. This aspect forms a critical distinction between ERM and RFL algorithms and underpins the success of the latter in achieving superior OOD performance. While existing RFL methods [10, 52] rely on iterative training on reweighted datasets to achieve this diversity, we propose a rigorous formulation of feature richness for featurizers, which naturally induce DOREEN. Further details and the complete proofs are provided in Sec. 7.

We start by establishing the foundational theoretical framework following the setups by Allen-Zhu and Li [3], Zhang and Bottou [51].

Definition 1 (Feature & Model [51])**.** *Let $(x, y) \sim P$ be a data point from the distribution P. We call feature a function $x \mapsto \phi(x) \in \mathbb{R}$. A deep learning model is denoted as $f = \omega^\top \Phi$. $\Phi = [\phi_1, \phi_2, ...\phi_n]^\top \in \mathbb{R}^n$ is a featurizer where ϕ_is are features and exploited with a linear classifier $\omega = [\omega_1, \omega_2, ...\omega_n]^\top \in \mathbb{R}^n$. For an input x, the output of the model f is $f(x) = \omega^\top \Phi(x) = \sum_{i=1}^{n} \omega_i \phi_i(x)$. The expected loss of a model f with a convex loss ℓ on data from distribution P is:*

$$\mathcal{L}_P(f) = \mathcal{L}_P(\omega, \Phi) = \mathbb{E}_{(x,y)\sim P}[\ell(\omega^\top \Phi(x), y)]. \quad (1)$$

And we make the following assumption about the optimality of the classifier based on the features.

Assumption 1 (Optimal classifier)**.** *Given a featurizer Φ, we assume the optimal classifier $\omega^* = \underset{\omega}{\operatorname{argmin}}\, \mathcal{L}_p(\omega, \Phi)$ is achievable by convex optimization methods. We use $\mathcal{L}_p^*(\Phi)$ to denote $\min_{\omega} \mathcal{L}_p(\omega, \Phi)$ for convenience.*

In this study, we focus on developing a richer featurizer, thus we will directly adopt the optimal classifier for clarity. Building on the concept of "multi-view" structure [3], we postulate the existence of multiple "informative" features within the given training data distribution. For instance, when identifying whether an animal is an elephant, we might extract the shape features to observe the trunk and large ear flaps. We can also examine the texture and color features to assess the distinctive, tough yet sensitive grey skin. We define "informative" as follows.

Definition 2 (Informative features)**.** *For a given training data distribution P_{tr}, there exists a set of underlying and informative features denoted as $S_{tr} = \{\phi_1^*, \phi_2^*, ..., \phi_t^*\}$ where $\mathcal{L}_{p_{tr}}^*(\phi_i^*) \leq \delta, \forall i$. δ is a constant that helps determine whether a feature is informative or not.*

Meanwhile, it is also natural to make the following assumption about the classifier's weights on uninformative or unseen features that have not appeared during training.

Assumption 2. *With the optimal classifier,* $f(x) = \omega^\top \Phi(x) = \sum_{\phi \in \mathcal{S}_{tr}} \omega_\phi \phi(x)$*, which means* $\omega_i = 0$ *if* $\phi_i \notin \mathcal{S}_{tr}$*. Intuitively, the classifier would not assign weights on uninformative features. Moreover, a* $\phi_j \in \mathcal{S}_{tr}$ *would get* $\omega_j \neq 0$*, while the later repeated* ϕ_j*s in the learned featurizer* Φ *would get zero weights.*

We then formalize a proxy metric for the informativity of a featurizer, measured by empirical risks.

Definition 3 (Set of informative & non-redundant features of a learned featurizer)**.** *Suppose a learned featurizer* $\Phi = [\tilde{\phi}_1, ..., \tilde{\phi}_k, \phi_1, \phi_1, \phi_2, \phi_2, ...\phi_m]^\top$ *where* $\forall i, \tilde{\phi}_i \notin \mathcal{S}_{tr}, \phi_i \in \mathcal{S}_{tr}$*. We then define* $S(\Phi) = \{\phi_1, \phi_2, ...\phi_m\}$*, representing the features extracted by* Φ *that are included in* $\mathcal{S}_{tr}$*. According to Assumption 1, we further say* $\mathcal{L}_p^*(\Phi) = \mathcal{L}_p^*(S(\Phi))$*.*

Under the aforementioned setup, for the informative features ($S_{tr} = \{\phi_1^*, \phi_2^*, ..., \phi_t^*\}$), there may exist some linear combinations $\phi_c = \sum_{i=1}^t \alpha_i * \phi_i^*$ such that $\mathcal{L}_{p_{tr}}^*(\phi_c) \leq \delta$ (see Sec. 7.1). If a feature extractor Φ_c learns ϕ_c, we let $S(\Phi_c) = \{\phi_i | \alpha_i \neq 0\}$. When a feature extractor learns their linear combination, we believe it has the potential to be distinguished into individual informative features. Then, based on the empirical observation above, we can establish a formal definition of feature richness.

Definition 4 (Feature richness)**.** *For two featurizers* Φ_1 *and* Φ_2 *learned on training data, we say* Φ_1 *extracts richer features than* Φ_2 *do iff* $S(\Phi_1) \supset S(\Phi_2)$*.*

Grounded in the theoretical framework that formally defines "richness", we analyze the existing RFL methods and identify potential shortcomings. We subsequently propose a more direct yet efficient approach to enhance Rich Feature Learning.

4. DOREEN Method and Analysis

DOREEN incorporates feature diversity into the loss function during training. This approach serves as a more effective means to foster the learning of rich features.

4.1. Analysis of existing RFL methods

We first conduct an analysis on the current RFL methods utilizing our theoretical framework. Due to the methodological similarities within the existing RFL methods in how they integrate features, we select FeAT [10] for our analysis, revealing that these methods might falter in extreme cases, resulting in a feature extractor Φ where $S(\Phi) = S(\Phi_{ERM})$, fail to integrate richer features.
In the k_{th} training round, FeAT obtains Φ^k by minimizing:

$$\mathcal{L}_{\text{FeAT}} = \max_{D_i^a \in G^a} \mathcal{L}_{D_i^a}(w_k^\top \Phi) + \lambda \sum_{D_i^r \in G^r} \mathcal{L}_{D_i^r}(w_i^\top \Phi) \quad (2)$$

where $G = \{G^r, G^a\}$ is a collection of datasets, divided into $2k$ subsets. The group for new feature augmentation is $G^a = \{D_i^a\}_{i=0}^{k-1}$, where D_j^a represents the subset of data points incorrectly predicted by the model in $(j-1)_{th}$ round and the initial augmentation set D_0^a corresponds to the entire training set D_{tr}. Conversely, $G^r = \{D_j^r\}_{j=0}^{k-1}$ comprises subsets of correctly predicted data for retaining features already learned and D_0^r is initially empty. The loss on subset D is defined as $\mathcal{L}_D(w^\top \Phi) = \frac{1}{|D|} \sum_{(x_i, y_i) \in D} \ell(\omega^\top \Phi(x_i), y_i)$. We denote by $\mathcal{L}_D^*(\Phi) = \min_\omega \mathcal{L}_D(\omega, \Phi)$ the loss achieved by the optimal classifier (see Assumption 1) on set D. In the first round, FeAT in fact conducts ERM training and $\Phi^1 = \Phi_{ERM}$.

Proposition 1 (FeAT fails with a small μ)**.** *If* Φ_{ERM} *satisfies* $\mathcal{L}_{D_{tr}}^*(\Phi_{ERM}) = \mu \leq \frac{\theta}{|D_{tr}|}$*, FeAT degrades to ERM. FeAT cannot learn* $\phi \in (\mathcal{S}_{tr} - S(\Phi_{ERM}))$*.*

The proof is given in Sec. 7.2. θ is introduced in Definition 5 due to the space limits. When the existing RFL methods incorporate features strongly correlated with the label, the augmentation set becomes negligible. This limitation hinders their ability to identify additional informative features, restricting the development of a richer featurizer according to Definition 4.

4.2. The DOREEN method

The richness formulation naturally induces DOREEN: an approach optimizing for both informativeness and diversity, involving two models with identical architecture and extendable to multiple models. This process minimizes a composite loss function comprising the ERM loss and diversity penalty, formally represented as: $\hat{\mathcal{L}}_{p_{tr}}(\Phi_k) = \mathcal{L}_{p_{tr}}(\Phi_k) + \mathcal{L}_{Div}(\Phi_k)$. The diversity penalty encompasses the inter-model part (the first term) and the intra-model part (the second term) as:

$$\mathcal{L}_{Div}(\Phi_k) = \alpha_k^1 * \sum_{i=1}^n \sum_{j=1}^n \mathbb{1}_{\Phi_{1i} = \Phi_{2j}} + \alpha_k^2 * \sum_{1 \leq i < j \leq n} \mathbb{1}_{\Phi_{ki} = \Phi_{kj}} \quad (3)$$

where $k \in \{1, 2\}$ is the index of different models, $\alpha_k^i, i \in \{1, 2\}$ are constant hyperparameters. The procedure of DOREEN is shown in Algorithm 1.

Due to the non-differentiability of the indicator function in Equation (3), we adopt the Determinantal Point Process (DPP) [28] as a regularizer [14, 49] to promote diversity in the outputs of the featurizer. It's worth noting that DOREEN is not tied to any specific diversity measure; rather, it serves as a flexible framework that bridges diversity concepts with rich feature learning. Users can freely tailor the choice of diversity measurement, such as Vendi Score or squared distance, to best suit their task and ensure optimal performance. DPP, a mathematical framework for modeling repulsion or diversity among a set of items, involves two key

Algorithm 1: The DOREEN Algorithm

Input : Training data D_{tr}; models $f_1 := \omega_1 \circ \Phi_1$, $f_2 := \omega_2 \circ \Phi_2$, ... , $f_k := \omega_k \circ \Phi_k$; training epochs e; hyperparameter α;

Randomly initialize $f_1, f_2, ..., f_k$;
for $i \leftarrow 1$ **to** e **do**
  Obtain $\mathcal{L}_{\text{ERM}}$ for $f_1, f_2, ..., f_k$;
  Randomly sample a subset $\mathcal{X}$ of D_{tr} and obtain $[\Phi_1(\mathcal{X}),\ \Phi_2(\mathcal{X}), ...,\ \Phi_k(\mathcal{X})]$;
  Compute $\mathcal{L}_{Div}$ for $\Phi_1, \Phi_2, ..., \Phi_k$;
  Update each f_i by minimizing $\mathcal{L}_{\text{ERM}}(f_i) + \alpha * \mathcal{L}_{Div}(\Phi_i)$;

Output: $\Phi = [\Phi_1\ \Phi_2\ ...\ \Phi_k]$

steps. First, it computes a kernel matrix L using a positive (semi-)definite kernel function to quantify the similarities among the items. Then, it calculates the determinant, or a related function, of L to gauge the overall diversity among these items. This incorporation of DPP into our loss function modification effectively ensures diversity in the feature extraction process.

Using Gaussian kernel function, we define the diversity loss based on DPP as:

$$\mathcal{L}_{\text{DPP}}(\Phi_1, \Phi_2) = \mathcal{D}et(\mathcal{K}([\Phi_1(\mathcal{X})\ \Phi_2(\mathcal{X})])) \tag{4}$$

$\mathcal{X} = \{x_i\}_{i=1}^m$ is a randomly sampled set of inputs, then $[\Phi_1(\mathcal{X})\ \Phi_2(\mathcal{X})]$ is a concatenation of outputs of Φ_1 and Φ_2 on $\mathcal{X}$ of size $m * 2n$, $\mathcal{K}(A = \{a_1, a_2 ... a_t\})$ where a_is are column vectors is a kernel matrix whose size is $t * t$ and $\mathcal{K}(A)_{(i,j)} = Sim(a_i, a_j)$ is the similarity between a_i and a_j measured by function $Sim(\cdot)$. $Sim(\cdot)$ must ensure that $\mathcal{K}$ is positive (semi-) definite [28].

Then we minimize the loss:

$$\hat{\mathcal{L}}_{p_{tr}}(\omega_k, \Phi_k) = \mathcal{L}_{\text{ERM}}(\omega_k, \Phi_k) + \alpha * \mathcal{L}_{\text{DPP}}(\Phi_1, \Phi_2) \tag{5}$$

where $k \in \{1, 2\}$ is the index of models, $\mathcal{L}_{\text{ERM}}(\omega, \Phi) = \frac{1}{|D_{tr}|} \sum_{(x_i, y_i) \in D_{tr}} \ell(\omega^{*\top} \Phi(x_i), y_i)$ is a standard ERM loss computed for two models respectively, $\mathcal{L}_{\text{DPP}}(\Phi_1, \Phi_2)$ is a shared diversity loss of both models. The analysis of additional computational overhead introduced by the diversity loss is provided in Sec. 7.5.

4.3. Improvement over ERM

In this subsection, we compare $\Phi_{\text{DOREEN}} = [\Phi_1 \quad \Phi_2]$ and Φ_{ERM}, where we let $\alpha_1^1 = \alpha_1^2 = 0$ for DOREEN and thus $\Phi_1 = \Phi_{ERM}$.

For ERM, if the current featurizer $\bar{\Phi}$ satisfies $\mathcal{L}^*_{p_{tr}}(S(\bar{\Phi})) = \mathcal{L}^*_{p_{tr}}(S(\bar{\Phi}) \cup \Phi_s) = \lambda$, $\forall \Phi_s \subseteq \mathcal{S}_{tr}$, then ERM cannot learn any $\phi \in (\mathcal{S}_{tr} - S(\bar{\Phi}))$. Intuitively, ERM rapidly acquires simple features that are effective on the training set. However, if these simple features exhibit a strong correlation with the labels within the training distribution, ERM may overlook additional, more complex yet beneficial features.

Proposition 2 (Inter-model diversity helps achieve feature richness)**.** *When* $\mathcal{L}^*_{p_{tr}}(S(\Phi_1)) = \mathcal{L}^*_{p_{tr}}(S(\Phi_1) \cup \Phi_s) = \lambda$ *for any* $\Phi_s \subseteq \mathcal{S}_{tr}$, Φ_2 *can learn* $\phi \in (\mathcal{S}_{tr} - S(\Phi_1))$ *if* α_2^1 *satisfies* $\alpha_2^1 > \delta - \lambda$, *then* $[\Phi_1\ \Phi_2]$ *is richer than* Φ_{ERM}.

The proof is provided in Sec. 7.3. When Φ_1 saturates the training objective and cannot incorporate additional informative features, inter-model diversity with weight $\alpha_2^1 > \delta - \lambda$ enables Φ_2 to capture richer features. In situations with correlations strongly correlated with the labels corresponding to a negligible λ, where conventional RFL methods may struggle as discussed in Sec. 4.1, DOREEN can effectively address these challenges by adjusting α_2^1. Furthermore, as discussed in Addepalli et al. [1] and supported by the empirical findings in Figure 6, ERM is hindered by feature replication. DOREEN can also address this issue through the intra-model diversity term (α_2^2), as shown in our analysis in Sec. 7.4.

5. Experiments

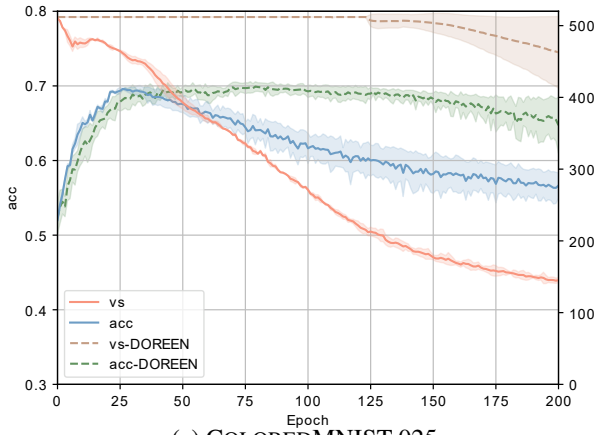
5.1. DOREEN vs. ERM

Experimental setups. We first conduct experiments to evaluate both feature diversity and OOD accuracy throughout the training process of DOREEN. The experimental setup mirrors that of Sec. 3.2.

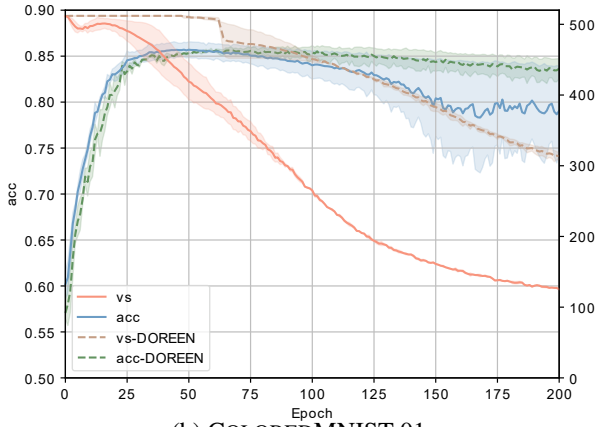
Results. The results in Figure 3 illustrate that DOREEN exhibits consistently higher feature diversity and ensures superior OOD performance. Notably, DOREEN not only matches or surpasses the peak OOD accuracy achieved by ERM but also maintains this performance across a broader range of training epochs. This stability provides a wider margin for effective model selection. In contrast, as shown in Figure 2 (b), ERM attains its peak OOD accuracy only within a narrow window, making it challenging to precisely pinpoint and capitalize on its optimal performance phase. To further analyze the impact of explicitly promoting diversity, we conducted an ablation study by training two ERM models with different random initializations, each with a hidden dimension of 256, and concatenating them to assess the feature diversity and OOD performance. As shown in Figure 3, the concatenated ERM model outperforms a single ERM model but still falls short of the stability exhibited by DOREEN. This deficiency arises from the limited diversity achieved solely through different random initializations, which does not fundamentally alter ERM's inherent characteristics. Throughout the training process, the feature diversity of the learned representations consistently decreases. This highlights the effectiveness and superiority of directly integrating feature diversity into DOREEN.

Table 1. OOD performance on ColoredMNIST datasets.

Algorithm	ColoredMNIST-025				ColoredMNIST-01				ColoredMNIST-sp		
	ERM	BONSAI	FeAT	DOREEN	ERM	BONSAI	FeAT	DOREEN	ERM	FeAT	DOREEN
ERM	12.40(±0.32)	11.21(±0.49)	17.27(±2.55)	**17.65**(±4.60)	73.75(±0.49)	70.95(±0.93)	76.05(±1.45)	**76.16**(±0.54)	10.00 (±0.35)	9.87(±0.43)	**18.62**(±2.11)
IRMv1	59.81(±4.46)	70.28(±0.72)	**70.57**(±0.68)	69.18(±0.87)	73.84(±0.56)	76.71(±4.10)	82.33(±1.71)	**82.55**(±1.88)	48.75(±2.60)	48.88(±2.67)	**56.21**(±3.21)
V-REX	65.96(±1.29)	70.31(±0.66)	**70.82**(±0.59)	69.61(±0.75)	81.20(±3.27)	82.61(±1.76)	84.70(±0.69)	**85.60**(±0.55)	49.01(±3.86)	49.66(±1.40)	**56.66**(±2.34)
IRMX	64.05(±0.88)	70.46(±0.42)	**70.78**(±0.61)	69.79(±0.64)	75.97(±0.88)	80.28(±1.62)	84.34(±0.97)	**85.53**(±0.97)	48.50(±2.80)	48.77(±2.05)	**52.30**(±1.69)
IB-IRM	59.81(±4.46)	70.28(±0.72)	**70.57**(±0.68)	69.36(±0.88)	73.84(±0.56)	76.71(±4.10)	82.33(±1.77)	**83.00**(±2.09)	48.62(±2.61)	48.70(±2.01)	**53.50**(±1.86)



(a) ColoredMNIST-025



(b) ColoredMNIST-01

Figure 3. Comparison between DOREEN and ensemble-based methods. The x-axis illustrates the evaluation epochs of the featurizer. The y-axis displays the OOD accuracy (acc) and Vendi Score (vs). The acc/vs-DOREEN: Training dynamics of DOREEN. The acc/vs: Training dynamics of the concatenation of two ERM models with different random initializations.

Table 2. OOD performance of different diversity techniques.

	Deep Ensemble	DiWA	WLD-Reg	[46]	DOREEN
ColoredMNIST-025	65.41 (±1.52)	59.01 (±3.97)	68.79 (±1.42)	65.30 (±3.09)	**69.61** (±0.75)
ColoredMNIST-01	82.47 (±1.05)	70.47 (±7.01)	85.06 (±0.92)	84.01 (±1.19)	**85.60** (±0.55)
ColoredMNIST-sp	51.46 (±1.88)	52.56 (±0.66)	55.26 (±3.12)	51.90 (±0.44)	**56.66** (±2.34)

5.2. A Controlled Study

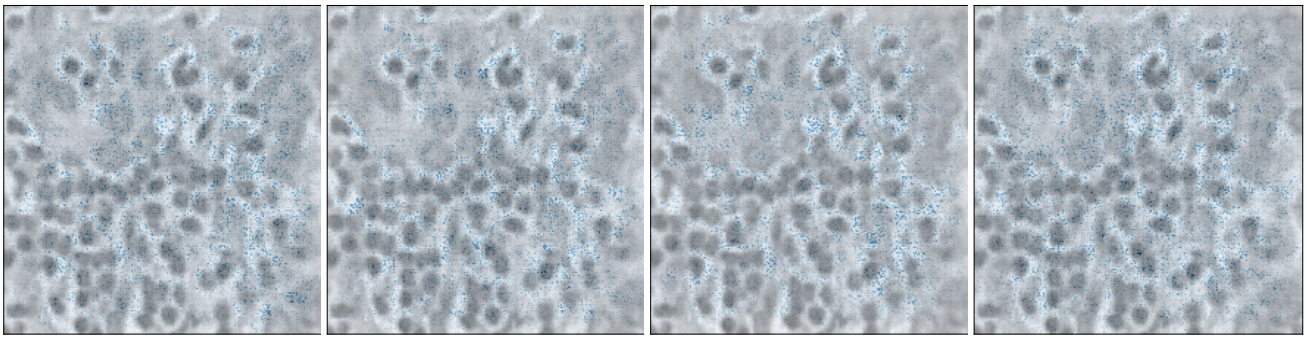
Experimental setups. We then conducted a controlled study on the ColoredMNIST datasets [4] to assess the feature learning capabilities of DOREEN under various conditions. In addition to the two previously mentioned ColoredMNIST datasets, we extended our experiments to ColoredMNIST-sp to represent scenarios with extreme spurious correlations and corroborate the assertions in Proposition 1. We compared the OOD performance of the features learned by DOREEN, against those acquired via ERM, BONSAI and FeAT. We employed a variety of representative OOD objectives, including IRMv1 [4], V-REX [27], IRMX [9] and IB-IRM [2] to evaluate the quality of the learned features. The detailed empirical settings are presented in Sec. 10.2.1.

Results. 1) Feature learning quality. As shown in Tab. 1, DOREEN consistently outperforms ERM across all three datasets. Compared to the two SOTA RFL methods, DOREEN performs slightly below FeAT and BONSAI on ColoredMNIST-025, but surpasses both on ColoredMNIST-01 and ColoredMNIST-sp, achieving the highest average performance overall. Notably, in scenarios with radical spurious correlations, BONSAI encounters issues due to an empty augmentation set. FeAT's performance aligns closely with that of ERM. This is consistent with our theoretical findings in Proposition 1 and confirms our concerns about the dependency of existing RFL methods on the quality of the augmentation set. In contrast, DOREEN shows marked improvements by leveraging the multi-view structure of the data, effectively learns richer features, demonstrating significantly enhanced performance.

2) Comparison with Other OOD Methods Incorporating Diversity Techniques. Since DivDis [31] and D-BAT [34] require access to unlabeled OOD data during training—a setting that can be expensive and impractical in many real-world applications [40, 43], we compare DOREEN with Deep Ensemble, DIWA, WLD-Reg, Teney et al. [46] listed in Figure 1 (b), focusing on the feature learning quality with V-REX serving as the OOD training objective. The empirical results, presented in Tab. 2, highlight the following: 1) By promoting diversity at the feature level, WLD-Reg and DOREEN significantly outperform the other methods. 2) DOREEN consistently outperforms WLD-Reg with fewer parameters and computational cost, emphasizing the advantages of inter-model diversity. 3) All diversity-promoting methods outperform ERM and FeAT on ColoredMNIST-sp. This supports our concerns regarding current RFL methods: they may struggle to develop a richer featurizer when faced with spurious correlations strongly tied to the labels, while promoting diversity can effectively address this limitation.

Table 3. OOD generalization performances on WILDS benchmark.

	CAMELYON17 (Avg. acc.)			FMOW (Worst acc.)			IWILDCAM (Macro F1)		
	DFR	V-REX	GroupDRO	DFR	V-REX	GroupDRO	DFR	V-REX	GroupDRO
ERM	95.14 (±1.96)	71.60 (±7.88)	76.09 (±6.46)	41.96 (±1.90)	33.06 (±0.46)	33.03 (±0.52)	23.15 (±0.24)	28.82 (±0.47)	28.51 (±0.58)
BONSAI	95.17 (±0.18)	76.39 (±5.32)	72.82 (±5.37)	43.26 (±0.82)	33.17 (±1.26)	33.12 (±1.20)	21.36 (±0.41)	25.81 (±0.42)	27.16 (±1.18)
FeAT	95.28 (±0.19)	75.12 (±6.55)	80.41 (±3.30)	43.54 (±1.26)	33.17 (±1.26)	34.04 (±0.70)	23.54 (±0.52)	29.48 (±1.94)	28.38 (±1.82)
DOREEN	**96.90** (±0.10)	**78.00** (±6.54)	**81.69** (±4.72)	**45.06** (±1.78)	**34.63** (±1.44)	**34.34** (±0.37)	**24.41** (±0.46)	**30.51** (±0.70)	**30.16** (±0.97)



(a) Camelyon17-ERM (b) Camelyon17-FeAT (c) Camelyon17-BONSAI (d) Camelyon17-DOREEN

Figure 4. Integrated Gradients visualization of models trained by ERM, BONSAI, FeAT and DOREEN on Camelyon17. The blue dots are the salient features. A deeper blue color denotes more salient features.

More empirical results and the corresponding analysis are listed in Sec. 10.2.2. Figure 7 provides additional validation that a model's OOD performance is strongly correlated with its feature diversity. In extreme scenarios, the feature diversity of the FeAT-trained featurizer shows minimal improvement over that of ERM, whereas DOREEN effectively learns richer features, demonstrating significantly enhanced performance. Figure 8 illustrates DOREEN's OOD performance under various penalty weights for diversity loss. This highlights DOREEN's resilience against fluctuations in hyperparameters. Notably, in scenarios with more pronounced spurious correlations, a higher penalty weight for diversity loss proves advantageous.

5.3. Feature Learning with Realistic Benchmarks

Finally, we compared DOREEN with ERM and the SOTA RFL methods in 3 real-world OOD generalization datasets curated by Koh et al. [25]: Camelyon17 [5], FMoW [12], and iWildCam [7], which contain complicated features and notable distribution shifts.

Empirical settings. The learned features are evaluated with V-REX and GroupDRO [41], two representative SOTA OOD objectives in WILDS. In addition to OOD objectives, we evaluate the learned features with Deep Feature Reweighting (DFR) [24]. DFR uses an additional OOD validation set where the spurious correlation does not hold to perform logistic regression based on the learned features. Thus, DFR serves as an effective proxy for evaluating the quality of learned invariant features [20]. Dataset and implementation details can be found in Sec. 10.3.

Empirical results. The results presented in Tab. 3 demonstrate that DOREEN consistently outperforms ERM and the two SOTA RFL methods across all three datasets and OOD objectives, validating its effectiveness in real-world scenarios. We also use Integrated Gradients [45] to assess the feature learning performance of different algorithms. The visualization shown in Figure 4 and Figure 9 demonstrates that DOREEN is able to learn more meaningful and diverse features.

6. Conclusion

In this study, we conducted a comprehensive investigation of RFL methods and ERM, highlighting the pivotal role of diversity in Rich Feature Learning. Our study not only provides a clear and formal definition of "rich features"—characterized by both diversity and informativeness—but also introduces DOREEN, a novel approach that explicitly promotes feature diversity at both the intra-model and inter-model levels to enhance Rich Feature Learning. Theoretically, we demonstrate that DOREEN effectively incorporates richer features than ERM. Furthermore, we identify and empirically validate that the existing RFL methods struggle under radical spurious correlations, whereas DOREEN efficiently handle such challenging scenarios. In our extensive experiments conducted across both controlled and realistic settings, the results consistently illustrate the superior performance of DOREEN. Serving as a bridge between diversity principles and rich feature learning rather than a method tied to a specific diversity optimizer, DOREEN thereby enables practitioners to choose the diversity notion best aligned with their data and constraints. We hope this framework will spur further work on diversity-aware feature learning and robust OOD generalization.

Acknowledgment

This work was supported in part by the Guangdong Basic and Applied Basic Research Foundation under Grant No. 2025A1515012968, in part by the Shenzhen Science and Technology Program under Grant No. JCYJ20240813113502004, in part by the National Natural Science Foundation of China under Grant No. 62001412, in part by Shenzhen Stability Science Program 2023, in part by the Guangdong Provincial Key Laboratory of Future Networks of Intelligence (Grant No. 2022B1212010001), and in part by the Shenzhen Key Lab of Crowd Intelligence Empowered Low-Carbon Energy Network (Grant No. ZDSYS20220606100601002), and in part by the School of Computing, National University of Singapore (grant no: A-0010308-00-00).

References

[1] Sravanti Addepalli, Anshul Nasery, Venkatesh Babu Radhakrishnan, Praneeth Netrapalli, and Prateek Jain. Feature reconstruction from outputs can mitigate simplicity bias in neural networks. In *The Eleventh International Conference on Learning Representations*, 2022. 6, 8

[2] Kartik Ahuja, Ethan Caballero, Dinghuai Zhang, Jean-Christophe Gagnon-Audet, Yoshua Bengio, Ioannis Mitliagkas, and Irina Rish. Invariance principle meets information bottleneck for out-of-distribution generalization. *Advances in Neural Information Processing Systems*, 34: 3438–3450, 2021. 1, 2, 7

[3] Zeyuan Allen-Zhu and Yuanzhi Li. Towards understanding ensemble, knowledge distillation and self-distillation in deep learning. *arXiv preprint arXiv:2012.09816*, 2020. 3, 4

[4] Martin Arjovsky, Léon Bottou, Ishaan Gulrajani, and David Lopez-Paz. Invariant risk minimization. *arXiv preprint arXiv:1907.02893*, 2019. 1, 2, 3, 7, 6

[5] Peter Bandi, Oscar Geessink, Quirine Manson, Marcory Van Dijk, Maschenka Balkenhol, Meyke Hermsen, Babak Ehteshami Bejnordi, Byungjae Lee, Kyunghyun Paeng, Aoxiao Zhong, et al. From detection of individual metastases to classification of lymph node status at the patient level: the camelyon17 challenge. *IEEE transactions on medical imaging*, 38(2):550–560, 2018. 8, 11

[6] Sara Beery, Grant Van Horn, and Pietro Perona. Recognition in terra incognita. In *Proceedings of the European conference on computer vision (ECCV)*, pages 456–473, 2018. 1

[7] Sara Beery, Arushi Agarwal, Elijah Cole, and Vighnesh Birodkar. The iwildcam 2021 competition dataset. *arXiv preprint arXiv:2105.03494*, 2021. 8, 11

[8] Harold Benoit, Liangze Jiang, Andrei Atanov, Ouguzhan Fatih Kar, Mattia Rigotti, and Amir Zamir. Unraveling the key components of ood generalization via diversification. *arXiv preprint arXiv:2312.16313*, 2023. 3, 5, 9

[9] Yongqiang Chen, Kaiwen Zhou, Yatao Bian, Binghui Xie, Kaili Ma, Yonggang Zhang, Han Yang, Bo Han, and James Cheng. Pareto invariant risk minimization. *arXiv preprint arXiv:2206.07766*, 2022. 2, 7

[10] Yongqiang Chen, Wei Huang, Kaiwen Zhou, Yatao Bian, Bo Han, and James Cheng. Understanding and improving feature learning for out-of-distribution generalization. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023. 1, 2, 3, 4, 5, 7, 8, 11

[11] Yongqiang Chen, Kaiwen Zhou, Yatao Bian, Binghui Xie, Bingzhe Wu, Yonggang Zhang, MA KAILI, Han Yang, Peilin Zhao, Bo Han, and James Cheng. Pareto invariant risk minimization: Towards mitigating the optimization dilemma in out-of-distribution generalization. In *The Eleventh International Conference on Learning Representations*, 2023. 1, 2

[12] Gordon Christie, Neil Fendley, James Wilson, and Ryan Mukherjee. Functional map of the world. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6172–6180, 2018. 8, 11

[13] Joseph Paul Cohen, Joseph D Viviano, Paul Bertin, Paul Morrison, Parsa Torabian, Matteo Guarrera, Matthew P Lungren, Akshay Chaudhari, Rupert Brooks, Mohammad Hashir, et al. Torchxrayvision: A library of chest x-ray datasets and models. In *International Conference on Medical Imaging with Deep Learning*, pages 231–249. PMLR, 2022. 3

[14] Mohamed Elfeki, Camille Couprie, Morgane Riviere, and Mohamed Elhoseiny. Gdpp: Learning diverse generations using determinantal point processes. In *International conference on machine learning*, pages 1774–1783. PMLR, 2019. 5

[15] Dan Friedman and Adji Bousso Dieng. The vendi score: A diversity evaluation metric for machine learning. *arXiv preprint arXiv:2210.02410*, 2022. 1, 4, 8

[16] Ishaan Gulrajani and David Lopez-Paz. In search of lost domain generalization. *arXiv preprint arXiv:2007.01434*, 2020. 2, 7

[17] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016. 11

[18] Gao Huang, Zhuang Liu, Laurens Van Der Maaten, and Kilian Q Weinberger. Densely connected convolutional networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4700–4708, 2017. 11

[19] Daksh Idnani, Vivek Madan, Naman Goyal, David J Schwab, and Shanmukha Ramakrishna Vedantam. Don't forget the nullspace! nullspace occupancy as a mechanism for out of distribution failure. In *The Eleventh International Conference on Learning Representations*, 2023. 5, 6

[20] Pavel Izmailov, Polina Kirichenko, Nate Gruver, and Andrew G Wilson. On feature learning in the presence of spurious correlations. *Advances in Neural Information Processing Systems*, 35:38516–38532, 2022. 8

[21] Yuanfeng Ji, Lu Zhang, Jiaxiang Wu, Bingzhe Wu, Long-Kai Huang, Tingyang Xu, Yu Rong, Lanqing Li, Jie Ren, Ding Xue, et al. Drugood: Out-of-distribution (ood) dataset curator and benchmark for ai-aided drug discovery–a focus on affinity prediction problems with noise annotations. *arXiv preprint arXiv:2201.09637*, 2022. 2

[22] Alex Kendall and Yarin Gal. What uncertainties do we need in bayesian deep learning for computer vision? *Advances in neural information processing systems*, 30, 2017. 3

[23] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014. 7

[24] Polina Kirichenko, Pavel Izmailov, and Andrew Gordon Wilson. Last layer re-training is sufficient for robustness to spurious correlations. *arXiv preprint arXiv:2204.02937*, 2022. 8

[25] Pang Wei Koh, Shiori Sagawa, Henrik Marklund, Sang Michael Xie, Marvin Zhang, Akshay Balsubramani, Weihua Hu, Michihiro Yasunaga, Richard Lanas Phillips, Irena Gao, et al. Wilds: A benchmark of in-the-wild distribution shifts. In *International Conference on Machine Learning*, pages 5637–5664. PMLR, 2021. 2, 8, 7, 11

[26] Masanori Koyama and Shoichiro Yamaguchi. Out-of-distribution generalization with maximal invariant predictor. 2020. 1, 2

[27] David Krueger, Ethan Caballero, Joern-Henrik Jacobsen, Amy Zhang, Jonathan Binas, Dinghuai Zhang, Remi Le Priol, and Aaron Courville. Out-of-distribution generalization via risk extrapolation (rex). In *International Conference on Machine Learning*, pages 5815–5826. PMLR, 2021. 1, 2, 4, 7

[28] Alex Kulesza, Ben Taskar, et al. Determinantal point processes for machine learning. *Foundations and Trends® in Machine Learning*, 5(2–3):123–286, 2012. 5, 6

[29] Firas Laakom, Jenni Raitoharju, Alexandros Iosifidis, and Moncef Gabbouj. Wld-reg: A data-dependent within-layer diversity regularizer. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 8421–8429, 2023. 2, 3

[30] Balaji Lakshminarayanan, Alexander Pritzel, and Charles Blundell. Simple and scalable predictive uncertainty estimation using deep ensembles. *Advances in neural information processing systems*, 30, 2017. 2

[31] Yoonho Lee, Huaxiu Yao, and Chelsea Finn. Diversify and disambiguate: Out-of-distribution robustness via disagreement. In *The Eleventh International Conference on Learning Representations*, 2022. 2, 3, 7, 5

[32] Vaishnavh Nagarajan, Anders Andreassen, and Behnam Neyshabur. Understanding the failure modes of out-of-distribution generalization. *arXiv preprint arXiv:2010.15775*, 2020. 5, 6

[33] Hongseok Namkoong and John C Duchi. Stochastic gradient methods for distributionally robust optimization with f-divergences. *Advances in neural information processing systems*, 29, 2016. 3

[34] Matteo Pagliardini, Martin Jaggi, Franccois Fleuret, and Sai Praneeth Karimireddy. Agree to disagree: Diversity through disagreement for better transferability. *arXiv preprint arXiv:2202.04414*, 2022. 2, 3, 7, 5

[35] Chau Pham and Bryan Plummer. Enhancing feature diversity boosts channel-adaptive vision transformers. *Advances in Neural Information Processing Systems*, 37:89782–89805, 2024. 5

[36] Alexandre Rame and Matthieu Cord. Dice: Diversity in deep ensembles via conditional redundancy adversarial estimation. *arXiv preprint arXiv:2101.05544*, 2021. 3

[37] Alexandre Rame, Corentin Dancette, and Matthieu Cord. Fishr: Invariant gradient variances for out-of-distribution generalization. In *International Conference on Machine Learning*, pages 18347–18377. PMLR, 2022. 1, 2

[38] Alexandre Rame, Matthieu Kirchmeyer, Thibaud Rahier, Alain Rakotomamonjy, Patrick Gallinari, and Matthieu Cord. Diverse weight averaging for out-of-distribution generalization. *Advances in Neural Information Processing Systems*, 35:10821–10836, 2022. 2, 3

[39] Elan Rosenfeld, Pradeep Ravikumar, and Andrej Risteski. Domain-adjusted regression or: Erm may already learn features sufficient for out-of-distribution generalization. *arXiv preprint arXiv:2202.06856*, 2022. 2

[40] Alexander Rubinstein, Luca Scimeca, Damien Teney, and Seong Joon Oh. Scalable ensemble diversification for ood generalization and detection. *arXiv preprint arXiv:2409.16797*, 2024. 7, 5

[41] Shiori Sagawa, Pang Wei Koh, Tatsunori B Hashimoto, and Percy Liang. Distributionally robust neural networks for group shifts: On the importance of regularization for worst-case generalization. *arXiv preprint arXiv:1911.08731*, 2019. 3, 8, 5, 6

[42] Luca Scimeca, Alexander Rubinstein, Armand Mihai Nicolicioiu, Damien Teney, and Yoshua Bengio. Leveraging diffusion disentangled representations to mitigate shortcuts in underspecified visual tasks. *arXiv preprint arXiv:2310.02230*, 2023. 5

[43] Luca Scimeca, Alexander Rubinstein, Damien Teney, Seong Joon Oh, Armand Mihai Nicolicioiu, and Yoshua Bengio. Mitigating shortcut learning with diffusion counterfactuals and diverse ensembles. *arXiv preprint arXiv:2311.16176*, 2023. 7, 5

[44] Yuge Shi, Jeffrey Seely, Philip HS Torr, N Siddharth, Awni Hannun, Nicolas Usunier, and Gabriel Synnaeve. Gradient matching for domain generalization. *arXiv preprint arXiv:2104.09937*, 2021. 1, 2

[45] Mukund Sundararajan, Ankur Taly, and Qiqi Yan. Axiomatic attribution for deep networks. In *Proceedings of the 34th International Conference on Machine Learning - Volume 70*, pages 3319–3328. JMLR.org, 2017. 8, 12

[46] Damien Teney, Ehsan Abbasnejad, Simon Lucey, and Anton Van den Hengel. Evading the simplicity bias: Training a diverse set of models discovers solutions with superior ood generalization. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 16761–16772, 2022. 2, 3, 7, 5, 8, 9

[47] Sebastian Tschiatschek, Rishabh K Iyer, Haochen Wei, and Jeff A Bilmes. Learning mixtures of submodular functions for image collection summarization. *Advances in neural information processing systems*, 27, 2014. 5

[48] Yu Wang, Junxian Mu, Pengfei Zhu, and Qinghua Hu. Exploring diverse representations for open set recognition. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 5731–5739, 2024. 5

[49] Pengtao Xie, Aarti Singh, and Eric P Xing. Uncorrelation and evenness: a new diversity-promoting regularizer. In *International Conference on Machine Learning*, pages 3811–3820. PMLR, 2017. 5

[50] Baosheng Yu, Meng Fang, Dacheng Tao, and Jie Yin. Submodular asymmetric feature selection in cascade object detection. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 2016. 5

[51] Jianyu Zhang and Léon Bottou. Learning useful representations for shifting tasks and distributions. In *International Conference on Machine Learning*, pages 40830–40850. PMLR, 2023. 4

[52] Jianyu Zhang, David Lopez-Paz, and Léon Bottou. Rich feature construction for the optimization-generalization dilemma. In *International Conference on Machine Learning*, pages 26397–26411. PMLR, 2022. 1, 2, 4, 7

[53] Tianren Zhang, Chujie Zhao, Guanyu Chen, Yizhou Jiang, and Feng Chen. Feature contamination: Neural networks learn uncorrelated features and fail to generalize. *arXiv preprint arXiv:2406.03345*, 2024. 5, 6