RICH FEATURE LEARNING VIA DIVERSIFICATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Rich Feature Learning (RFL) aims to extract all beneficial features from the training distribution and has demonstrated significant efficacy in Out-of-Distribution (OOD) generalization. Despite its success, a precise and comprehensive definition of "richness" remains elusive. Through an in-depth analysis of RFL algorithms and empirical risk minimization (ERM), the standard OOD baseline, we identify feature diversity as the key differentiator driving RFL's superior OOD performance. Building on this insight, we formally define rich features as those that exhibit both high 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. 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.

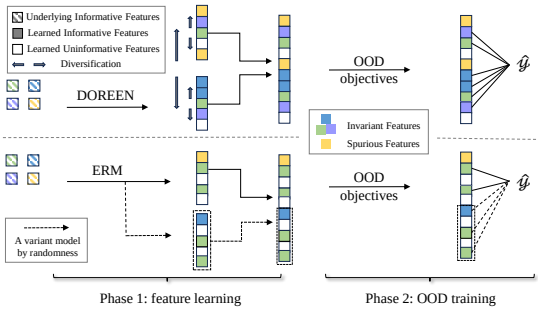
1 INTRODUCTION

The significant performance degradation of models trained with ERM on OOD data is commonly attributed to learning spurious features (Arjovsky et al., 2019; Beery et al., 2018). To address this challenge, there has been a surge of efforts in developing OOD objectives to regularize ERM feature learning (Arjovsky et al., 2019; Ahuja et al., 2021; Krueger et al., 2021; Shi et al., 2021; Koyama & Yamaguchi, 2020; Rame et al., 2022a; Chen et al., 2023b). However, such regularization can disrupt the standard ERM feature learning process, introducing substantial optimization dilemma (Zhang et al., 2022; Chen et al., 2023a).

To overcome these optimization difficulties, the concept of Rich Feature Learning (RFL) was introduced (Zhang et al., 2022; Chen et al., 2023a). 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 sets "rich features" apart from those learned via ERM. Specifically, we assess the diversity of the features learned by ERM together with two SOTA RFL algorithms: BONSAI (also called RFC) (Zhang et al., 2022) and FeAT (Chen et al., 2023a) using Vendi Score (Friedman & Dieng, 2022) and examine the corresponding OOD performance. In Section 3 we first compare the diversity and OOD performance of the features learned by ERM with those of RFL methods. The results indicate that ERM-trained featurizers exhibit lower feature diversity and inferior OOD performance compared to RFL methods. Then 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 highlight that feature diversity is the primary factor distinguishing RFL from ERM. Consequently, we propose a formal definition of rich features as those that are both *diverse* and *informative*. Based on this theoretical foundation, in Section 4, we first demonstrate that the existing SOTA RFL methods can fail in scenarios with strong spurious correlations. Then we propose a simple yet powerful RFL algorithm, DOREEN (**D**iversity-f**O**unded **R**ich f**E**ature l**E**arni**N**g). We show that, by embedding diversity directly into the learning process, DOREEN can integrate richer features than ERM and effectively addresses scenarios where existing RFL methods struggle.



(a)

Algorithm	RFL	Diversity Usage	Implementation	Scope
ERM	✗	✗	N/A	N/A
BONSAI (Zhang et al., 2022)	✓	✗	N/A	N/A
FeAT (Chen et al., 2023a)	✓	✗	N/A	N/A
DiWA (Rame et al., 2022b)	✗	✓	implicit	whole model
Deep Ensemble (Lakshminarayanan et al., 2017)	✗	✓	implicit	whole model
Teney et al. (2022)	✗	✓	explicit	classifier
DivDis (Lee et al., 2022)	✗	✓	explicit	classfier
WLD-Reg Laakom et al. (2023)	✗	✓	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, fostering robust representations well-suited for OOD training. (b) Overview of diversity-based algorithms for OOD generalization: Algorithms are classified as "explicitly" if they include a diversity penalty. Only BONSAI, FeAT, and DOREEN are RFL methods.

DOREEN trains multiple models with identical 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). Extensive experiments (Section 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) Novel algorithm and theoretical analysis: Based on the observation, we propose a formal definition for rich features and introduce a novel RFL algorithm named DOREEN. We provide theoretical evidence for its ability to incorporate richer features than ERM and address 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 (Arjovsky et al., 2019; Chen et al., 2022; Ahuja et al., 2021; Krueger et al., 2021; Shi et al., 2021; Koyama & Yamaguchi, 2020; Rame et al., 2022a). However, the optimization challenges posed by these objectives often surpass the complexities of ERM (Chen et al., 2023b) and introduce heavy disturbance to the ERM feature learning (Zhang et al., 2022), leading to empirical observations by Zhang et al. (2022) 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. Contrasting with 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 focuses on improving OOD performance by developing feature representations that are broader and more comprehensive than those learned by ERM. BONSAI (Zhang et al., 2022) constructs rich representations by iteratively learning new features from incorrectly predicted subsets (augmentation sets) while retaining previously learned features that 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 (Chen et al., 2023a) efficiently learn rich representations by optimizing a combined loss that includes ERM loss on the retention sets and DRO loss (Namkoong & Duchi, 2016) 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.

Leveraging Diversity for Enhanced OOD Generalization. Allen-Zhu & Li (2020) argue that numerous parameter values can adequately explain observations within finite training data which they call "multi-view" and emphasize the importance of capturing this diversity for robust OOD performance (Kendall & Gal, 2017). Multi-view structures are common in real-world data—for example, lions can be identified by either their manes or faces. Capturing these diverse views facilitates robust predictions, even in test sets that include female lions without manes. There has been a surge of efforts aimed at leveraging diversity to enhance OOD generalization. Rame et al. (2022b) propose weight averaging across multiple models to encourage diversity. However, this approach primarily relies on the randomness of initialization to generate model variance, failing to explicitly promote diversity among them, which may lead to redundancy (Rame & Cord, 2021). Teney et al. (2022); Lee et al. (2022) instead opt to capture diverse features by training multiple classifiers and ensuring they produce distinct predictions. Our approach stands apart by directly encouraging diversity among feature extractors, enabling the learning of a wider range of features that generalize to new distributions. Additionally, while Laakom et al. (2023) 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.

Comparison to previous works. A detailed comparison of DOREEN with relevant algorithms is presented in Figure 1(b). As an RFL method, DOREEN explicitly introduces both inter-model and intra-model diversity into feature extractors, enabling the development of richer features.

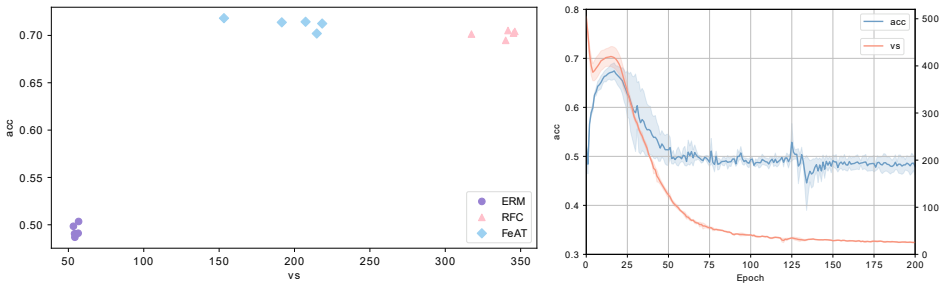
3 Motivating Studies and Featurizer Richness

To explore the distinctive aspects of "rich features" compared to those learned by ERM, we initiated a series of experiments focused on feature diversity utilizing the ColoredMNIST dataset (Arjovsky et al., 2019) (denoted as ColoredMNIST-025). We also encompass 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 4. Detailed information about the ColoredMNIST dataset and other configurations of these experiments are provided in Appendix C.1.

We begin our analysis by comparing ERM and two RFL algorithms: BONSAI (also known as RFC) and FeAT. To assess feature diversity, we utilize the Vendi Score, as conceptualized by Friedman & Dieng (2022). 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 (Krueger et al., 2021) to train a classifier atop it to evaluate the OOD performance on test data, in that V-REX is a SOTA OOD objective and 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.

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 5 further validate the point.

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



(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.

in achieving superior OOD performance. While existing RFL methods Zhang et al. (2022); Chen et al. (2023a) 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. Due to space constraints, the theoretical foundations, including the necessary definitions and assumptions, are provided in Appendix A.

Definition 1 (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 then 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.

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 Chen et al. (2023a) 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.

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 can not learn $\phi \in (\mathcal{S}_{tr} - S(\Phi_{ERM}))$.*

Where θ is introduced in Definition 5 and the detailed proof is provided in Appendix B.2 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 1.

The DOREEN method. The richness formulation naturally induce DOREEN: an approach optimizing on both informativeness and diversity that involves two models of identical structure 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:

$$\mathcal{L}_{Div}(\Phi_k) = \alpha_k^1 * \sum_{i=1}^{n}\sum_{j=1}^{n} \mathbb{1}_{\Phi_{1_i}=\Phi_{2_j}} + \alpha_k^2 * \sum_{1\leq i<j\leq n} \mathbb{1}_{\Phi_{k_i}=\Phi_{k_j}} \tag{1}$$

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

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

Due to the non-differentiability of the indicator function in Equation (1), we adopt the Determinantal Point Process (DPP) Kulesza et al. (2012) as a regularizer Elfeki et al. (2019); Xie et al. (2017) to promote diversity in the outputs of the featurizer. Alternatives such as the Vendi Score or various pairwise similarity metrics can also be utilized to assess diversity within the framework of DOREEN. Users can flexibly tailor the choice of diversity measurement to align with the specific requirements of their task to ensuring optimal performance. 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:

$$\mathcal{L}_{\text{DPP}}(\Phi_1, \Phi_2) = \mathcal{D}et(\mathcal{K}([\Phi_1(\mathcal{X})\ \Phi_2(\mathcal{X})])) \tag{2}$$

$\mathcal{X} = \{x_i\}_{i=1}^m$ 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 Kulesza et al. (2012).

Then we minimize the loss:

$$\hat{\mathcal{L}}_{p_{tr}}(\omega_k, \Phi_k) = \mathcal{L}_{\text{ERM}}(\omega_k, \Phi_k) + \alpha * \mathcal{L}_{\text{DPP}}(\Phi_1, \Phi_2) \tag{3}$$

where k is the index of different 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 Appendix B.5.

Improvement over ERM. 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 $\Phi_1 = \Phi_{ERM}$. The detailed proof is provided in Appendix B.3.

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}.

Thus, DOREEN is capable of capturing richer features by setting a sufficiently large α_2^1. In situations with correlations strongly correlated with the labels corresponding to a negligible λ, where current RFL methods may struggle, as discussed earlier, DOREEN can effectively address these challenges by adjusting α_2^1 appropriately. Furthermore, as discussed in Addepalli et al. (2022) and demonstrated by the empirical findings in Figure 5, ERM is hindered by feature replication. DOREEN can also address this issue with a relatively large α_2^2 (intra-model diversity) in our analysis shown in Appendix B.4.

5 Experiments

We first mirror the experiments in Section 3 to assess the feature diversity and OOD performance during the training process of DOREEN. The results in Figure 3 illustrate that DOREEN exhibits

Table 1: Results on two 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)	**85.80**(±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)

consistently higher feature diversity and ensures superior OOD performance. Detailed empirical settings and further analysis are provided in Appendix C.2.

A controlled study. We then conducted a controlled study on the ColoredMNIST dataset (Arjovsky et al., 2019) 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 characterized by $\varepsilon_{tr} = \{(0.1, 0), (0.1, 0)\}$ 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. Detailed empirical settings are listed in Appendix C.3. As presented in Table 1, DOREEN demonstrates a notable improvement over ERM in all three datasets. When compared to the two SOTA RFL algorithms, DOREEN exhibits a slightly lower performance on ColoredMNIST-025 but surpasses both FeAT and BONSAI on ColoredMNIST-01 and ColoredMNIST-sp, achieving the highest overall average performance across the three datasets. 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. We also conduct comparison between DOREEN and other OOD methods incorporating diversity techniques: Deep Ensemble, DIWA, WLD-Reg. The results in Table 5 illustrate that directly encouraging diversity among feature extractors help to yield richer representations and stronger OOD performance. Detailed empirical settings, more empirical results and comprehensive analysis and are listed in Appendix C.3.2.

Table 2: OOD performances on the WILDS benchmark.

Dataset	OOD Training	ERM	BONSAI	FeAT	DOREEN
CAMELYON17	V-REX	71.60 (±7.88)	76.39 (±5.32)	75.12 (±6.55)	**76.84**(±5.88)
	GroupDRO	76.09 (±6.46)	72.82 (±5.37)	80.41 (±3.30)	**81.64**(±4.38)
	DFR	95.14 (±1.96)	95.17 (±0.18)	95.28 (±0.19)	**96.90**(±0.10)
FMOW	V-REX	33.06 (±0.46)	33.17 (±1.26)	34.00 (±0.71)	**34.60** (±0.62)
	GroupDRO	33.03 (±0.52)	33.12 (±1.20)	34.04 (±0.70)	**35.21** (±0.30)
	DFR	41.96 (±1.90)	43.26 (±0.82)	43.54 (±1.26)	**45.06**(±1.78)

Feature learning with realistic benchmarks. Finally, we compared DOREEN with ERM, BONSAI and FeAT in 2 real-world OOD generalization datasets: Camelyon17 Bandi et al. (2018) and FMoW Christie et al. (2018) that contain complicated features and notable distribution shifts. More details about the datasets and empirical settings can be found in Appendix C.4. The results, shown in Table 2, demonstrate that DOREEN consistently outperforms ERM and the two SOTA RFL methods across both datasets and all three OOD objectives, validating the effectiveness of DOREEN in real-world scenarios. We further use Integrated Gradients Sundararajan et al. (2017) to assess the feature learning performance of different algorithms. The visualization shown in Figure 8 demonstrate that DOREEN is able to learn more meaningful and diverse features than ERM, BONSAI and FeAT.

6 Conclusion

In this study, we have undertaken a thorough investigation into the RFL methods and ERM, highlighting the critical role of diversity in Rich Feature Learning. Our study not only presents a clear and formal definition of "rich features" – characterized as diverse and informative – but also introduces DOREEN, a novel approach designed to enhance feature diversity and thereby facilitate Rich Feature Learning. Theoretically, we demonstrate that DOREEN effectively incorporate richer features than ERM. Furthermore, we identify and empirically validate that the existing RFL methods falter when confronted with radical spurious correlations while 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.

REFERENCES

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.

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.

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

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

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.

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

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.

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*, 2023a.

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*, 2023b.

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*, pp. 6172–6180, 2018.

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

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

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

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.

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.

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

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.

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*, pp. 5637–5664. PMLR, 2021.

Masanori Koyama and Shoichiro Yamaguchi. Out-of-distribution generalization with maximal invariant predictor. 2020.

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*, pp. 5815–5826. PMLR, 2021.

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

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*, volume 37, pp. 8421–8429, 2023.

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

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.

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

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

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

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, 2022b.

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.

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.

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

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*, pp. 16761–16772, 2022.

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

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

Jianyu Zhang, David Lopez-Paz, and Léon Bottou. Rich feature construction for the optimization-generalization dilemma. In *International Conference on Machine Learning*, pp. 26397–26411. PMLR, 2022.

A THEORETICAL FOUNDATION OF FEATURE RICHNESS

We start by establishing the foundational theoretical framework following the setups by Allen-Zhu & Li (2020); Zhang & Bottou (2023).

Definition 2 (Feature & Model Zhang & Bottou (2023))**.** *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 a 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:*

$$\mathcal{L}_P(f) = \mathcal{L}_P(\omega, \Phi) = \mathbb{E}_{(x,y)\sim P}[\ell(\omega^\top \Phi(x), y)]. \tag{4}$$

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 $\omega^* = \operatorname{argmin}_\omega \mathcal{L}_p(\omega, \Phi)$ 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" Allen-Zhu & Li (2020), 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 but sensitive grey skin. We define "informative" as follows.

Definition 3 (Informative features)**.** *For a given training data distribution P_{tr}, there exists an 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 helping to distinguish whether a feature is informative or not.*

Meanwhile, it is also natural to establish the following assumption about the classifier weights onto the uninformative and 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_js 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 4 (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 $\mathcal{L}_p^*(\Phi) = \mathcal{L}_p^*(S(\Phi))$.*

With the aforementioned setup, for the informative features ($S_{tr} = \{\phi_1^*, \phi_2^*, ..., \phi_t^*\}$), there may exist some linear combinations $\phi_c = \sum_{i=1}^t \alpha_i * \phi_i^*$ satisfies $\mathcal{L}_{p_{tr}}^*(\phi_c) \leq \delta$. (The complete proof is available in Appendix B.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 have 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.

B PROOFS

We first introduce the notations used for theoretical analysis in Table 3

Table 3: Notations for key concepts involved in this paper

Symbols	Definitions
$x \in \mathbb{R}^m$	A single input
$\mathcal{X} = \{(x_i, y_i)\}_{i=1}^N$	A set of inputs
$\phi : \mathbb{R}^m \to \mathbb{R}$	A single feature
n	the hidden dimension
$\Phi = [\phi_1, \phi_2, ...\phi_n]^\top : \mathbb{R}^m \to \mathbb{R}^n$	A featurizer
Φ_i	The i_{th} feature in featurizer Φ
$\omega = [\omega_1, \omega_2, ...\omega_n]^\top \in \mathbb{R}^n$	A linear classifier
$f = \omega \circ \Phi : \mathbb{R}^m \to \mathbb{R}$	A predictor (model)
P	A certain data distribution
P_{tr}	The training data distribution
$\ell : \mathbb{R} \times \mathbb{R} \to \mathbb{R}$	A convex loss function
$\mathcal{L}_P(\cdot)$	The expected loss of a model on data distribution P
ω^*	The optimal classifier
$\mathcal{L}_p^*(\Phi)$	The expected loss of featurizer Φ on distribution P with the optimal classifier
$Sim(,) \in \mathbb{R}$	the similarity between two features
$\mathcal{S}_{tr}$	The set of informative features under the training distribution
$S(\Phi)$	The set of features extracted by featurizer Φ that are included in $\mathcal{S}_{tr}$
$\mathcal{L}_{Div}$	The diversity penalty
α_k^1	The penalty weight of inter-model diversity penalty of the k_{th} model
α_k^2	The penalty weight of intra-model diversity penalty of the k_{th} model
$\mathcal{L}_{\text{DPP}}$	The diversity penalty based on DPP

B.1 THE LINEAR COMBINATION OF THE INFORMATIVE FEATURES.

We first assume the utilized loss $\ell(\cdot)$ is convex and define the addition as well as scalar multiplication on features as follows:

- Addition: for features ϕ_1, ϕ_2, $(\phi_1 + \phi_2)(x) = \phi_1(x) + \phi_2(x), \forall x \in \mathbb{R}^m$.
- Scalar multiplication: for feature ϕ, $(\lambda \cdot \phi)(x) = \lambda \cdot \phi(x), \forall x \in \mathbb{R}^m$.

Then, for the informative features ($S_{tr} = \{\phi_1^*, \phi_2^*, ..., \phi_t^*\}$), there may exist some linear combinations $\phi_c = \sum_{i=1}^t \alpha_i * \phi_i^*$ satisfies $\mathcal{L}_{p_{tr}}^*(\phi_c) \leq \delta$. If a feature extractor Φ_c learns ϕ_c, we let $S(\Phi_c) = \{\phi_i | \alpha_i \neq 0\}$. That is, when a feature extractor learns their linear combination, we believe it have the potential to be distinguished into individual informative features.

Proof. We proved the case with two informative features and the result can be easily extended into the case with multiple informative features. For $\phi_1, \phi_2 \in S_{tr}$:

- Scalar multiplication: assume that $\operatorname{argmin}_\omega \mathcal{L}_{P_{tr}}(\omega, \phi) = \omega^*$, then for any real number λ:

$$\mathcal{L}_{p_{tr}}^*(\lambda * \phi) = \mathcal{L}_{P_{tr}}(\frac{\omega^*}{\lambda}, \phi) = \delta$$

- Addition: assume that $\operatorname{argmin}_\omega \mathcal{L}_{P_{tr}}(\omega, \phi_1) = \omega_1$, $\operatorname{argmin}_\omega \mathcal{L}_{P_{tr}}(\omega, \phi_2) = \omega_2$. we have:

$$\begin{aligned}\mathcal{L}_{p_{tr}}^*(\phi_1 + \phi_2) &\leq \mathcal{L}_{P_{tr}}(\frac{\omega_1\omega_2}{\omega_1 + \omega_2}, \phi_1 + \phi_2) = \mathbb{E}_{(x,y)\sim P_{tr}}[\ell(\frac{\omega_1\omega_2}{\omega_1 + \omega_2}(\phi_1 + \phi_2)(x), y)] \\ &= \mathbb{E}_{(x,y)\sim P_{tr}}[\ell(\frac{\omega_2}{\omega_1 + \omega_2}\omega_1 * \phi_1(x) + \frac{\omega_1}{\omega_1 + \omega_2}\omega_2 * \phi_2(x), y)]\end{aligned}$$

If ω_1, ω_2 satisfy that $0 \leq \frac{\omega_2}{\omega_1+\omega_2} \leq 1$, since $\ell(\cdot)$ is a convex function, we will have:

$$\begin{aligned}\mathcal{L}^*_{p_{tr}}(\phi_1+\phi_2) &\leq \frac{\omega_2}{\omega_1+\omega_2}\mathbb{E}_{(x,y)\sim P_{tr}}[\ell(\omega_1 * \phi_1(x), y)] \\ &+ \frac{\omega_1}{\omega_1+\omega_2}\mathbb{E}_{(x,y)\sim P_{tr}}[\ell(\omega_2 * \phi_2(x), y)] \\ &= \frac{\omega_2}{\omega_1+\omega_2} * \mathcal{L}^*_{p_{tr}}(\phi_1) + \frac{\omega_1}{\omega_1+\omega_2} * \mathcal{L}^*_{p_{tr}}(\phi_2) \\ &\leq \frac{\omega_2}{\omega_1+\omega_2} * \delta + \frac{\omega_1}{\omega_1+\omega_2} * \delta = \delta\end{aligned}$$

□

Before we delve into the analysis of DOREEN, we first present two lemmas about inter-model and intra-model diversity to help with the following proof. Suppose a learned featurizer $\Phi = [\tilde{\phi_1}, ..., \tilde{\phi_k}, \phi_1, \phi_1, \phi_2, \phi_2, ...\phi_m]^\top$ is of length n, we let $\Phi(x)$ be indexed as $\Phi(x)_1 = \tilde{\phi_1}(x), \Phi(x)_n = \phi_m(x)$. With two feature extractors Φ_1, Φ_2, we divide the diversity penalty $\mathcal{L}_{Div}(\Phi_k)$ into the inter-model part $div(\Phi_1, \Phi_2) = \alpha_k^1 * \sum_{i=1}^n \sum_{j=1}^n \mathbb{1}_{\Phi_{1i}=\Phi_{2j}}$ and intra-model part $div(\Phi_k) = \alpha_k^2 * \sum_{1\leq i<j\leq n} \mathbb{1}_{\Phi_{ki}=\Phi_{kj}}$.

Lemma B.1. *$div(\Phi_1, \Phi_2) = div(\Phi_1, \Phi_2/\Phi_{2t}) + div(\Phi_1, \Phi_{2t})$. Here Φ_2/Φ_{2t} is a featurizer of dimension n-1, without the t-th feature in Φ_2, and we slightly abuse Φ_{2t} in $div(\Phi_1, \Phi_{2t})$ to mean a featurizer of dimension 1 with only the t-th feature in Φ_2.*

Proof.

$$\begin{aligned}div(\Phi_1, \Phi_2) &= \sum_{1\leq i\leq n}\sum_{1\leq j\leq n} \mathbb{1}_{\Phi_{1i}=\Phi_{2j}} \\ &= \sum_{1\leq i\leq n}\sum_{1\leq j\leq n} (\mathbb{1}_{j=t} + \mathbb{1}_{j\neq t})\mathbb{1}_{\Phi_{1i}=\Phi_{2j}} \\ &= \sum_{1\leq i\leq n}\sum_{1\leq j\leq n} \mathbb{1}_{j\neq t}\mathbb{1}_{\Phi_{1i}=\Phi_{2j}} + \sum_{1\leq i\leq n}\sum_{1\leq j\leq n} \mathbb{1}_{j=t}\mathbb{1}_{\Phi_{1i}=\Phi_{2j}} \\ &= \sum_{1\leq i\leq n}\sum_{1\leq j\leq n, j\neq t} \mathbb{1}_{\Phi_{1i}=\Phi_{2j}} + \sum_{1\leq i\leq n} \mathbb{1}_{\Phi_{1i}=\Phi_{2j}} \\ &= div(\Phi_1, \Phi_2/\Phi_{2t}) + div(\Phi_1, \Phi_{2t}).\end{aligned}$$

□

Lemma B.2. *Given one featurizer Φ, the diversity penalty $div(\Phi) = div(\Phi/\Phi_t) + div(\Phi/\Phi_t, \Phi_t)$. Here Φ/Φ_t is a featurizer of dimension n-1, without the t-th feature in Φ, and we slightly abuse the second Φ_t in $div(\Phi/\Phi_t, \Phi_t)$ to mean a featurizer of dimension 1 with only the t-th feature of Φ.*

Proof.

$$\begin{aligned}div(\Phi) &= \sum_{1\leq i<j\leq n} \mathbb{1}_{\Phi_i=\Phi_j} \\ &= \sum_{1\leq i<j\leq n} (\mathbb{1}_{j=t} + \mathbb{1}_{j\neq t})\mathbb{1}_{\Phi_i=\Phi_j} \\ &= \sum_{1\leq i<j\leq n} \mathbb{1}_{j=t}\mathbb{1}_{\Phi_i=\Phi_j} + \sum_{1\leq i<j\leq n} (\mathbb{1}_{i=t} + \mathbb{1}_{i\neq t})\mathbb{1}_{j\neq t}\mathbb{1}_{\Phi_i=\Phi_j} \\ &= \sum_{1\leq i<t} \mathbb{1}_{\Phi_i=\Phi_t} + \sum_{1\leq i<j\leq n} \mathbb{1}_{i=t}\mathbb{1}_{j\neq t}\mathbb{1}_{\Phi_i=\Phi_j} + \sum_{1\leq i<j\leq n} \mathbb{1}_{i\neq t}\mathbb{1}_{j\neq t}\mathbb{1}_{\Phi_i=\Phi_j} \\ &= \sum_{1\leq i<t} \mathbb{1}_{\Phi_i=\Phi_t} + \sum_{t<j\leq n} \mathbb{1}_{\Phi_t=\Phi_j} + \sum_{1\leq i<j\leq n} \mathbb{1}_{i\neq t}\mathbb{1}_{j\neq t}\mathbb{1}_{\Phi_i=\Phi_j} \\ &= div(\Phi/\Phi_t, \Phi_t) + div(\Phi/\Phi_t).\end{aligned}$$

□

B.2 Proof of Proposition 1

In the k_{th} training round, FeAT obtain Φ^k by minimizing:

$$\mathcal{L}_{\text{FeAT}} = \max_{D_i^a \in G^a} \mathcal{L}_{D_i^a}(w_k^\top \Phi) + \lambda \sum_{D_i^r \in G^r} \mathcal{L}_{D_i^r}(w_i^\top \Phi) \tag{5}$$

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 use $\mathcal{L}_D^*(\Phi) = \min_\omega \mathcal{L}_D(\omega, \Phi)$ to denote the loss with the optimal classifier (Assumption 1) in set D. In the first round, FeAT in fact conduct ERM training and $\Phi^1 = \Phi_{ERM}$.

Before we start the proof, we need another definition as follows.

Definition 5 (Correct prediction)**.** *For a single data point x_i and its corresponding label y_i, we say the model $f = \omega^\top \Phi$ correctly predict on x_i if $\ell(\omega^\top \Phi(x_i), y_i) \leq \theta$.*

Then if Φ_{ERM} satisfies $\mathcal{L}_{D_{tr}}^*(\Phi_{ERM}) = \mu \leq \frac{\theta}{|D_{tr}|}$, FeAT degrades to ERM and can not learn $\phi \in (\mathcal{S}_{tr} - S(\Phi_{ERM}))$.

Proof. Assume there exists data points incorrectly predicted by Φ_{ERM} together with the best classifier ω^*, then $\mathcal{L}_{D_{tr}}^*(\Phi_{ERM}) = \frac{1}{|D_{tr}|} \sum_{(x_i,y_i)\in D_{tr}} \ell({\omega^*}^\top \Phi(x_i), y_i) = \mu$ would be larger than $\frac{\theta}{|D_{tr}|}$, contradictory to $\mu \leq \frac{\theta}{|D_{tr}|}$, thus we have $D_1^a = \varnothing$ as D_j^a represents the subset of data points that are incorrectly predicted by the model in the $(j-1)_{th}$ round.

When $D_1^a = \varnothing$, in the second round, Equation (5) degrades to :

$$\mathcal{L}_{\text{FeAT}} = (1+\lambda)\mathcal{L}_{D_{tr}}(\Phi) = (1+\lambda)\mathcal{L}_{\text{ERM}}$$

Then FeAT fails to learn richer featurizer than ERM in the later rounds whatever rounds it runs. □

B.3 Proof of Proposition 2

For ERM, If the current featurizer $\bar{\Phi}$ satisfying $\mathcal{L}_{p_{tr}}^*(S(\bar{\Phi})) = \mathcal{L}_{p_{tr}}^*(S(\bar{\Phi}) \cup \Phi_s) = \lambda, \forall \Phi_s \subseteq \mathcal{S}_{tr}$, then ERM can not learn $\phi \in (\mathcal{S}_{tr} - S(\bar{\Phi}))$. 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 neglect to learn additional, more complex but beneficial features.

(Inter-model diversity helps incorporate new informative features) 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}.

Proof. For simplicity, we assume now $S(\Phi_2) = \varnothing$, which means current Φ_2 is filled with uninformative features. we have:

- $\mathcal{L}_{p_{tr}}^*(\tilde{\phi}) \geq \lambda, \forall \tilde{\phi} \in S(\Phi_1), \lambda \leq \delta$.

$$\begin{aligned}
\delta \geq \mathcal{L}_{p_{tr}}^*(\tilde{\phi}) &= \min_{\omega_1} \mathcal{L}_{p_{tr}}(\omega_1, \tilde{\phi}) \\
&\geq \min_{\omega=\omega_1,\ldots,\omega_m} \mathcal{L}_{p_{tr}}(\omega, S(\Phi_1) = \{\tilde{\phi}, \tilde{\phi}_1, \ldots, \tilde{\phi}_{m-1}\}) \\
&= \mathcal{L}_{p_{tr}}^*(S(\Phi_1)) = \lambda.
\end{aligned}$$

- $\mathcal{L}_{p_{tr}}^*(\phi) \leq \delta, \forall \phi \in (\mathcal{S}_{tr} - S(\Phi_1))$, according to Definition 3 .

Then for Φ_2, the loss by Equation (1) is:

- when it learns $\tilde{\phi} \in S(\Phi_1)$

$$\begin{aligned}\hat{\mathcal{L}}_{p_{tr}}(\Phi_2) &= \mathcal{L}^*_{p_{tr}}(\Phi_2) + \alpha_2^1 \cdot div(\Phi_1, \Phi_2) + \alpha_2^2 \cdot div(\Phi_2) \\ &= \mathcal{L}^*_{p_{tr}}(\tilde{\phi}) + \alpha_2^1 \cdot \big(div(\Phi_1, \Phi_2/\tilde{\phi}) + div(\Phi_1, \tilde{\phi})\big) + \alpha_2^2 \cdot \big(div(\Phi_2/\tilde{\phi}) + div(\Phi_2/\tilde{\phi}, \tilde{\phi})\big) \\ &= \mathcal{L}^*_{p_{tr}}(\tilde{\phi}) + \alpha_2^1 \cdot (div(\Phi_1, \Phi_2/\tilde{\phi}) + 1) + \alpha_2^2 \cdot (div(\Phi_2/\tilde{\phi}) + 0) \\ &= \mathcal{L}^*_{p_{tr}}(\tilde{\phi}) + \alpha_2^1 + \alpha_2^1 \cdot div(\Phi_1, \Phi_2/\tilde{\phi}) + \alpha_2^2 \cdot div(\Phi_2/\tilde{\phi})\end{aligned}$$

  Let $\alpha_2^1 * div(\Phi_1, \Phi_2/\tilde{\phi}) + \alpha_2^2 * div(\Phi_2/\tilde{\phi}) = \eta$, we have:

$$\hat{\mathcal{L}}_{p_{tr}}(\Phi_2) \geq \lambda + \alpha_2^1 + \eta$$

- when it learns $\phi \in (\mathcal{S}_{tr} - S(\Phi_1))$

$$\begin{aligned}\hat{\mathcal{L}}_{p_{tr}}(\Phi_2) &= \mathcal{L}^*_{p_{tr}}(\Phi_2) + \alpha_2^1 * div(\Phi_1, \Phi_2) + \alpha_2^2 * div(\Phi_2) \\ &= \mathcal{L}^*_{p_{tr}}(\phi) + \alpha_2^1 * (div(\Phi_1, \Phi_2/\phi) + div(\Phi_1, \phi)) + \alpha_2^2 * (div(\Phi_2/\phi) + div(\Phi_2/\phi, \phi)) \\ &= \mathcal{L}^*_{p_{tr}}(\phi) + \alpha_2^1 * (div(\Phi_1, \Phi_2/\phi) + 0) + \alpha_2^2 * (div(\Phi_2/\phi) + 0) \\ &= \mathcal{L}^*_{p_{tr}}(\phi) + \alpha_2^1 * div(\Phi_1, \Phi_2/\phi) + \alpha_2^2 * div(\Phi_2/\phi) \\ &= \mathcal{L}^*_{p_{tr}}(\phi) + \eta \\ &\leq \delta + \eta\end{aligned}$$

- when it learns $\hat{\phi} \notin \mathcal{S}_{tr}$

$$\begin{aligned}\hat{\mathcal{L}}_{p_{tr}}(\Phi_2) &= \mathcal{L}^*_{p_{tr}}(\Phi_2) + \alpha_2^1 * div(\Phi_1, \Phi_2) + \alpha_2^2 * div(\Phi_2) \\ &= \mathcal{L}^*_{p_{tr}}(\hat{\phi}) + \alpha_2^1 * (div(\Phi_1, \Phi_2/\hat{\phi}) + div(\Phi_1, \hat{\phi})) + \alpha_2^2 * (div(\Phi_2/\hat{\phi}) + div(\Phi_2/\hat{\phi}, \hat{\phi})) \\ &\geq \mathcal{L}^*_{p_{tr}}(\hat{\phi}) + \alpha_2^1 * (div(\Phi_1, \Phi_2/\hat{\phi}) + 0) + \alpha_2^2 * (div(\Phi_2/\hat{\phi}) + 0) \\ &\geq \mathcal{L}^*_{p_{tr}}(\hat{\phi}) + \alpha_2^1 * div(\Phi_1, \Phi_2/\hat{\phi}) + \alpha_2^2 * div(\Phi_2/\hat{\phi}) \\ &\geq \mathcal{L}^*_{p_{tr}}(\hat{\phi}) + \eta \\ &> \delta + \eta\end{aligned}$$

Then Φ_2 will learn $\phi \in (\mathcal{S}_{tr} - S(\Phi_1))$ if $\delta + \eta < \lambda + \alpha_2^1 + \eta$, i.e., $\alpha_2^1 > \delta - \lambda \geq 0$. □

B.4 Intra-model diversity helps minigate feature replication

We use $\mathcal{R}(\Phi, \phi)$ to denote the times that ϕ replicates in Φ, $\mathcal{R}(\Phi, \phi) \geq 1$ for $\phi \in \Phi$ and $\mathcal{R}(\Phi, \phi) = 0$ for $\phi \notin \Phi$. Appendix B.4 depicts that the frequency of replication across different features does not exhibit significant variation.

Proposition 3 (Intra-model diversity helps minigate feature replication)**.** $\max_{\phi} \mathcal{R}(\Phi_2, \phi) - \min_{\phi} \mathcal{R}(\Phi_2, \phi) \leq 2$ *if* $\alpha_2^2 > n * \alpha_2^1$.

Proof. Assume that there are feature replication in Φ_1 and $\max_{\phi} \mathcal{R}(\Phi_1, \phi) = q \leq n$, there exists two features ϕ and $\tilde{\phi}$ such that $\mathcal{R}(\Phi, \phi) - \mathcal{R}(\Phi, \tilde{\phi}) = k \geq 2$ in the current featurizer Φ, now we look at another featurizer $\hat{\Phi}$ which is the same as Φ but substitute one ϕ with $\tilde{\phi}$.

$$\begin{aligned}\hat{\mathcal{L}}_{p_{tr}}(\hat{\Phi}) &= \mathcal{L}^*_{p_{tr}}(S(\hat{\Phi})) + \alpha_2^1 * div(\Phi_1, \hat{\Phi}) + \alpha_2^2 * div(\hat{\Phi}) \\ &= \mathcal{L}^*_{p_{tr}}(S(\hat{\Phi})) + \alpha_2^1 * (div(\Phi_1, \hat{\Phi}/\tilde{\phi}) + div(\Phi_1, \tilde{\phi})) + \alpha_2^2 * (div(\hat{\Phi}/\tilde{\phi}) + div(\hat{\Phi}/\tilde{\phi}, /\tilde{\phi})) \\ &= \alpha_2^1 * div(\Phi_1, \tilde{\phi}) + \alpha_2^2 * div(\hat{\Phi}/\tilde{\phi}, /\tilde{\phi}) + \eta_1\end{aligned}$$

where $\eta_1 = \mathcal{L}^*_{p_{tr}}(S(\hat{\Phi})) + \alpha_2^1 * div(\Phi_1, \hat{\Phi}/\tilde{\phi}) + \alpha_2^2 * div(\hat{\Phi}/\tilde{\phi})$.

$$\begin{aligned}\hat{\mathcal{L}}_{p_{tr}}(\Phi) &= \mathcal{L}^*_{p_{tr}}(S(\Phi)) + \alpha_2^1 * div(\Phi_1, \Phi) + \alpha_2^2 * div(\Phi) \\ &= \mathcal{L}^*_{p_{tr}}(S(\Phi)) + \alpha_2^1 * (div(\Phi_1, \Phi/\phi) + div(\Phi_1, \phi)) + \alpha_2^2 * (div(\Phi/\phi) + div(\Phi/\phi, /\phi)) \\ &= \alpha_2^1 * div(\Phi_1, \phi) + \alpha_2^2 * div(\Phi/\phi, /\phi) + \eta_2\end{aligned}$$

where $\eta_2 = \mathcal{L}^*_{p_{tr}}(S(\Phi)) + \alpha_2^1 * div(\Phi_1, \Phi/\phi) + \alpha_2^2 * div(\Phi/\phi)$.

Then

$$\hat{\mathcal{L}}_{p_{tr}}(\hat{\Phi}) - \hat{\mathcal{L}}_{p_{tr}}(\Phi) = \alpha_2^1 * (div(\Phi_1, \tilde{\phi}) - div(\Phi_1, \phi)) + \alpha_2^2 * (div(\hat{\Phi}/\tilde{\phi}, /\tilde{\phi}) - div(\Phi/\phi, /\phi)) + \eta_1 - \eta_2$$

We have $\eta_1 - \eta_2 = 0$ since:

- $\mathcal{L}^*_{p_{tr}}(S(\Phi)) = \mathcal{L}^*_{p_{tr}}(S(\hat{\Phi}))$ according to Assumption 2
- $\Phi/\phi = \hat{\Phi}/\tilde{\phi}$

Thus

$$\begin{aligned}\hat{\mathcal{L}}_{p_{tr}}(\hat{\Phi}) - \hat{\mathcal{L}}_{p_{tr}}(\Phi) &= \alpha_2^1 * (div(\Phi_1, \tilde{\phi}) - div(\Phi_1, \phi)) + \alpha_2^2 * (div(\hat{\Phi}/\tilde{\phi}, /\tilde{\phi}) - div(\Phi/\phi, /\phi)) \\ &\leq \alpha_2^1 * q + \alpha_2^2 * (1 - k)\end{aligned}$$

let $\hat{\mathcal{L}}_{p_{tr}}(\hat{\Phi}) - \hat{\mathcal{L}}_{p_{tr}}(\Phi) < 0$, we get $\alpha_2^2 > \frac{q * \alpha_2^1}{k-1}$, since $q \leq n$,$k \geq 2$, we finally get $\alpha_2^2 > n * \alpha_2^1$.

□

B.5 Additional computational overhead of the diversity loss

For simplicity, let's consider the computational overhead of a MLP with N hidden layers. Assume the input layer has dimension n_0, the i^{th} hidden layer has dimension n_i, and the output layer has dimension n_{N+1}. During a single training iteration over m examples, the time complexity for calculating the ERM loss is $O(m \sum_{i=1}^{N+1} n_i n_{i-1})$.

Directly leveraging the outputs from the penultimate layer across a tiny training subset of size k, the diversity loss calculation breaks down into two stages: constructing the similarity matrix, which is $O(n_N^2 k)$, and calculating the determinant of this matrix by computing and multiplying its eigenvalues, which is $O(n_N^3)$. If we choose k to be roughly equal to n_N, then the complexity for diversity loss computation simplifies to $O(n_N^3)$. In practical scenarios where m is much larger than n_N, the incremental computational cost of the diversity loss is almost negligible.

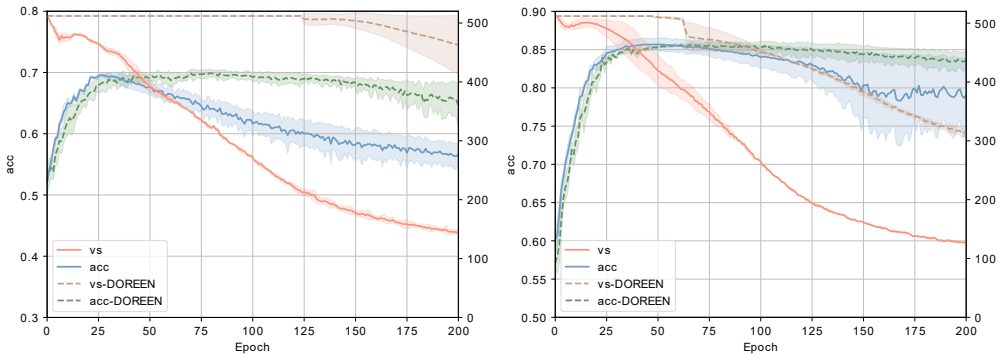
C Detailed Experiments

In this section, we provide more details and the implementation, evaluation and hyperparameter setups in complementary to the experiments in Section 3 and Section 5.

C.1 More details about the experiments in the Motivating studies

Vendi Score. 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 $\mathbf{K}/n$.

Datasets. We conducted experiments on the ColoredMNIST dataset (Arjovsky et al., 2019), including the original version where $\varepsilon_{tr} = \{(0.25, 0.1), (0.25, 0.2)\}$(denoted as as ColoredMNIST-025) and a modified ColoredMNIST (denoted as ColoredMNIST-01) with $\varepsilon_{tr} = \{(0.1, 0.2), (0.1, 0.25)\}$. The ColoredMNIST-025 is generated as follows: first, assign a preliminary binary label $\tilde{y}$ to the image based on the digit: $\tilde{y} = 0$ for digits 0-4 and $\tilde{y} = 1$ for 5-9. Second, obtain the final label y by flipping $\tilde{y}$ with probability 0.25. Third, sample the color id z by flipping y with probability p^e (0.1/0.2). The distinction between the two versions of the ColoredMNIST dataset lies in the feature-label correlation: spurious (ColoredMNIST-025) or invariant(ColoredMNIST-01) features are better correlated with labels.



(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.

Architecture and optimization. We use a 4-Layer MLP with a hidden dimension of 512 as the backbone model for all methods, where we take the first 3 layers as the featurizer and the last layer as the classifier, following the common practice (Gulrajani & Lopez-Paz, 2020; Koh et al., 2021). For the optimization of the models, we use the Adam (Kingma & Ba, 2014) optimizer with a learning rate of $1e-3$ and a weight decay of $1e-3$. For the model pretrained by ERM, BONSAI and FeAT, we use V-REX (Krueger et al., 2021) to train the classifier and report results.

Implementation of feature learning and OOD training methods. For the common feature learning protocol with ERM and two SOTA RFL methods: BONSAI (Zhang et al., 2022), FeAT (Chen et al., 2023a), our implementation follows (Chen et al., 2023a). For experiments in Section 3, we use V-REX (Krueger et al., 2021) as the OOD objective to apply the OOD regularization and adopt the implementations from (Zhang et al., 2022).

C.2 More details about the experiments assessing the feature diversity and OOD performance of DOREEN

Experimental setups. We conduct experiments to evaluate the feature diversity and OOD accuracy during the training process of DOREEN. The experimental setup mirrors that of the experiments in Section 3 to track the evolution of the feature diversity and OOD performance throughout the training process of a featurizer trained with ERM, but we train two models, each with a dimension of 256, and then concatenate them to form a 512-dimensional model in DOREEN.

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 of 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, Figure 2 (b) reveals that ERM reaches its peak OOD accuracy only during a brief phase, making it challenging to precisely pinpoint and leverage its optimal performance window.

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 decrease. This highlights the effectiveness and superiority of directly integrating feature diversity into DOREEN.

Table 4: Number of epochs in each round of various feature learning algorithms.

ColoredMNIST-025	Round-1	Round-2	Round-3	Syn. Round	ColoredMNIST-01	Round-1	Round-2	Round-3	Syn. Round	ColoredMNIST-sp	Round-1	Round-2
ERM	150	-	-	-	ERM	500	-	-	-	ERM	150	-
BONSAI	50	150	-	500	BONSAI	150	400	-	500	BONSAI	×	×
FeAT	150	150	-	-	FeAT	150	150	150	-	FeAT	150	150
DOREEN	300	-	-	-	DOREEN	500	-	-	-	DOREEN	150	-

C.3 More details about the experiments in the controlled study

In this section we show the detailed empirical settings and more results in the controlled study on ColoredMNIST. We conducted all the experiments utilizing NVIDIA GeForce RTX 3090.

C.3.1 Details about the experimental settings

As for dataset, in addition to the original version where $\varepsilon_{tr} = \{(0.25, 0.1), (0.25, 0.2)\}$(denoted as as ColoredMNIST-025) and a modified ColoredMNIST (denoted as ColoredMNIST-01) with $\varepsilon_{tr} = \{(0.1, 0.2), (0.1, 0.25)\}$, we further utilize ColoredMNIST-sp, characterized by $\varepsilon_{tr} = \{(0.1, 0), (0.1, 0)\}$, to address scenarios with extreme spurious correlations and corroborate the assertions in Proposition 1. For architecture and optimization, the settings are the same as that of Appendix C.1, except that we use a hidden dimension of 256 as in (Chen et al., 2023a) to obtain a fair comparison.

Implementation of feature learning and OOD training methods. For the common feature learning protocol with ERM and two SOTA RFL methods: BONSAI (Zhang et al., 2022), FeAT (Chen et al., 2023a), our implementation follows (Chen et al., 2023a). For OOD objectives, we adopt the implementations from (Zhang et al., 2022) for IRMv1 (Arjovsky et al., 2019),V-REX (Krueger et al., 2021) and IB-IRM (Ahuja et al., 2021); and the implementations from (Chen et al., 2023a) for IRMX (Chen et al., 2022).

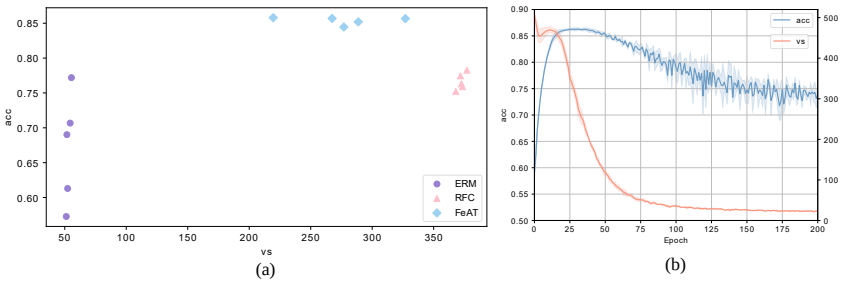
Evaluation of feature learning methods. For the sake of fairness in comparison, by default, we train all feature learning methods by the same number of epochs and rounds (if applicable). We strictly follow the recommended setups provided by (Zhang et al., 2022) for BONSAI and (Chen et al., 2023a) for FeAT. The settings of training rounds and epochs in each round are shown in Table 4. For the experiments on ColoredMNIST-025 and ColoredMNIST-01, we reported findings from Chen et al. (2023a) for ERM, BONSAI and FeAT, using the same empirical settings for DOREEN to ensure comparability. For experiments on ColoredMNIST-sp, classifiers can not find the causal features with any OOD objectives when only trained with the training set, so we instead train the classifiers with access to data from the test distribution (excluding the exact test set data), while keeping all other settings consistent.

Implementation of OOD Methods Incorporating Diversity Techniques. While Deep Ensemble, DIWA, and DOREEN leverage multiple models, WLD-Reg promotes diversity within a single feature extractor. For fairness, we configured Deep Ensemble, DIWA, and DOREEN with two models (each with a hidden dimension of 256) and WLD-Reg with a single model (hidden dimension of 512). We used the DPP loss as the diversity metric for WLD-Reg, aligning it with DOREEN. Other experimental parameters, such as training epochs and learning rates, were kept consistent across all methods, with V-REX serving as the OOD training objective.

C.3.2 More empirical results

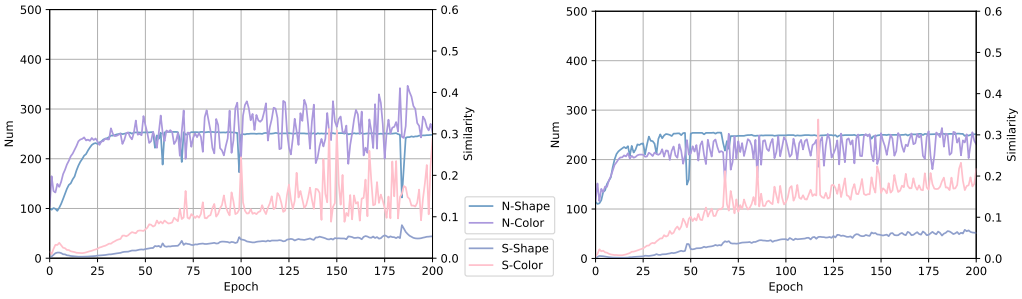
Comparison between ERM and RFL methods together with ERM training dynamics on ColoredMNIST-01. Figure 4(a) reveals that features learned through ERM exhibit significantly lower diversity compared to those obtained via RFC and FeAT. This, in turn, leads to a notably lower OOD accuracy for ERM when these features are applied for inference in contrast to the performance achieved by RFC and FeAT. Moreover, ERM-trained featurizers transiently possess high feature diversity and show promising OOD performance but this diversity diminishes as training advances, resulting in a parallel decrease in OOD performance as shown in Figure 4(b).

Shifts of Features during ERM Training Process We further sought to analyze the features developed during ERM training. We began by calculating the linear correlation between the outputs of



(a) (b)

Figure 4: The empirical results on ColoredMNIST-01. 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.



(a) ColoredMNIST-025 (b) ColoredMNIST-01

Figure 5: ERM training dynamics. The x-axis illustrates the evaluation epochs of the featurizer. The y-axis shows the number of extracted features (N-) and their intrinsic similarities (S-). At each evaluation epoch, we categorize each feature as related to shape, color, or as uninformative by calculating its linear correlation with the two label types, applying a predefined threshold value for categorization. Following this, we compute the similarities within the identified groups of shape and color features, respectively.

the penultimate layer of the ERM-trained model and the shape or color labels, identifying dimensions that strongly correlate (surpassing a set threshold) with either shape (shape features) or color (color features). We then assessed the feature similarity within these identified groups using the Average Pairwise Similarity Score (APSS) with the exponential similarity function(Friedman & Dieng, 2022). The results, detailed in Figure 5, reveal an initial increase in the number of features, which quickly reaches a plateau. Meanwhile, the similarity within these features continues to intensify. This pattern echoes the findings on feature diversity and OOD accuracy presented in Figure 2(b) and Figure 4(b), collectively indicating that representations with greater feature diversity yield improved OOD robustness. Furthermore, these empirical findings are in line with the Feature Replication Hypothesis by Addepalli et al. (2022), suggesting that simplicity bias drives the repeated learning of simpler features at the expense of more complex ones.

Feature diversity & OOD accuracy on ColoredMNIST-sp. We further plot the empirical results of ERM, FeAT and DOREEN on ColoredMNIST-sp. The results in Figure 6 further demonstrate that the OOD performance of a model is strongly correlated to the diversity of its featurizer. Moreover, in this extreme scenario, the feature diversity of FeAT-trained featurizer gains few improvement over that of ERM, leading to a almost the same or even worse OOD performance compared to ERM, while DOREEN effectively learns richer features and show apparently powerful performance.

Hyper parameter tuning. In our experiments utilizing Determinantal Point Processes (DPP) to apply a diversity penalty, two critical hyperparameters require tuning: sigma and penalty weight.

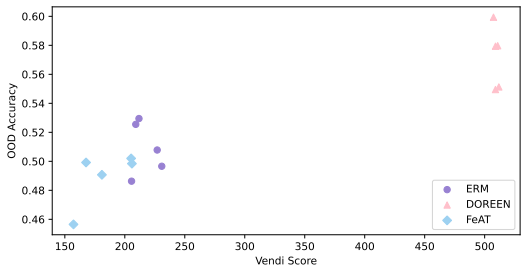
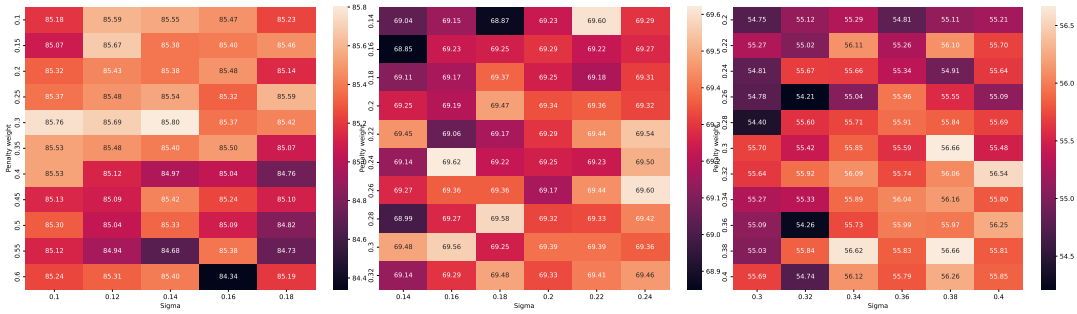


Figure 6: The feature diversity (x-axis) and OOD performance (y-axis) of featurizers trained with ERM, FeAT and DOREEN over five different random seeds. Post featurizer training, we measure the feature diversity using the Vendi Score, freeze the featurizer, and subsequently train a classifier using V-REX for OOD performance assessment.



(a) ColoredMNIST-01 (b) ColoredMNIST-025 (c) ColoredMNIST-sp

Figure 7: OOD performance corresponding to various combinations of sigma and penalty weight. In our experiments, we calibrated the range of our grid search to roughly make that the weighted diversity loss close to the ERM loss, maintaining a balance between informativeness and diversity.

Sigma serves as the hyperparameter for the DPP's kernel function, specifically an Gaussian kernel in our case, which measures the similarity between item pairs. A larger sigma value tends to increase kernel values, even for items that are relatively distant in feature space. Too large a sigma can lead to a more uniform kernel matrix, potentially diminishing the DPP's effectiveness in fostering diversity. On the other hand, too small a sigma might ignore useful similarities. Intuitively, when examining a specific group of items, a smaller sigma value typically indicates a higher degree of diversity among them. Conversely, a larger sigma value suggests that the items are less diverse. The key is finding a balance that captures meaningful similarities while still promoting diversity. The penalty weight, on the other hand, strikes a balance between diversity and informativeness. If set too high, it may lead to the selection of diverse yet uninformative features. In our experiments, we calibrated the range of our grid search to roughly make that the weighted diversity loss close to the ERM loss, maintaining a balance between informativeness and diversity. The out-of-distribution (OOD) performance, corresponding to various combinations of sigma and penalty weight, is illustrated in Figure 7. This visualization provides an understanding of how these hyperparameter adjustments impact the overall effectiveness of our approach. On one hand, This highlights DOREEN's resilience against fluctuations in hyperparameters. On the other hand, in scenarios with more pronounced spurious correlations where identifying diverse features is more crucial, a higher value for sigma and penalty weight of diversity loss proves advantageous.

Comparison with other OOD methods incorporating diversity techniques. The empirical results, presented in Table 5, highlight the following: 1) Relying solely on random initialization to foster diversity among models (e.g., in Deep Ensemble and DiWA) can lead to redundancy and instability. 2) DOREEN outperforms WLD-Reg with fewer parameters and the same computational cost, emphasizing the advantages of inter-model diversity. 3) All four 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 effectively addresses this limitation.

Table 5: Results of different diversity techniques.

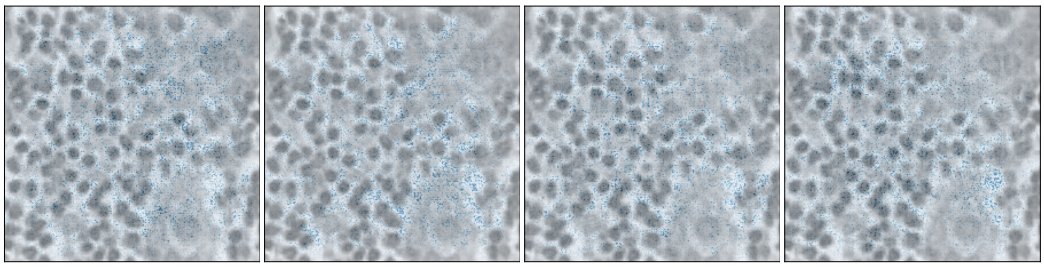
Method	ColoredMNIST-025	ColoredMNIST-01	ColoredMNIST-sp
Deep Ensemble	65.41(±1.52)	82.47(±1.05)	51.46(±1.88)
DiWA	59.01(±3.97)	70.47(±7.01)	52.56(±0.66)
WLD-Reg	68.79(±1.42)	85.06(±0.92)	55.26(±3.12)
DOREEN	**69.61**(±0.75)	**85.80**(±0.55)	**56.66**(±2.34)

Table 6: Hyperparameter setups of feature learning algorithms for the experiments on WILDS

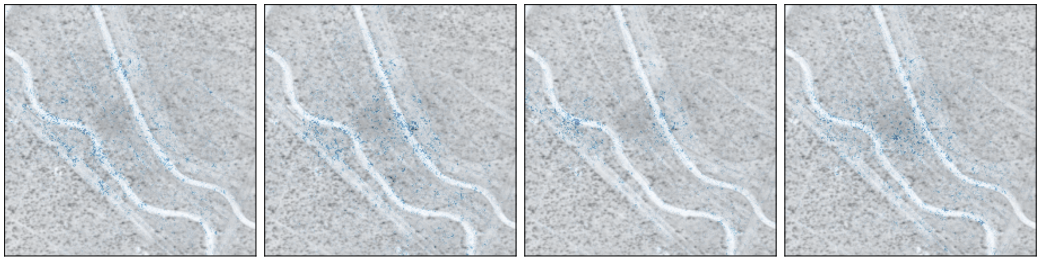
Dataset	Overall steps	Approx. epochs	Num. of rounds	Steps per round	Penalty weight
CAMELYON17	10000	10	2	5000	0.3
FMOW	9600	4	2	4800	0.1

C.4 More details about the WILDS experiments

In this section, we delve into further details about the WILDS datasets utilized in our experiments and describe our evaluation methodologies. Our investigation into feature learning performance under realistic conditions led us to choose two particularly challenging datasets from the WILDS benchmark. (Koh et al., 2021): Camelyon17 (Bandi et al., 2018) and FMoW (Christie et al., 2018). These datasets are characterized by a range of realistic distribution shifts, including domain distribution shifts, subpopulation shifts, and their combinations. Camelyon17 provides 450,000 lymph-node scans from 5 hospitals. The task is to take the input of 96×96 medical images to predict whether there exists a tumor tissue in the image. The domains d refers to the index of the hospital where the image was taken. The training data are sampled from the first 3 hospitals where the OOD validation and test data are sampled from the 4-th and 5-th hospital, respectively. FMoW provides satellite images from 16 years and 5 regions. The task in FMoW is to classify the images into 62 classes of building or land use categories. The domain is split according to the year that the satellite image was collected, as well as the regions in the image which could be Africa, America, Asia, Europe or Oceania. Distribution shifts could happen across different years and regions. The training data contains data collected before 2013, while the validation data contains images collected within 2013 to 2015, and the test data contains images collected after 2015. Comprehen-



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



(e) FMoW-ERM (f) FMoW-BONSAI (g) FMoW-FeAT (h) FMoW-DOREEN

Figure 8: Integrated Gradients visualization of models trained by ERM, BONSAI, FeAT and DOREEN on Camelyon17 and FMoW. The blue dots are the salient features. A deeper blue color denotes more salient features.

Table 7: General hyperparameter settings for the experiments on WILDS

Dataset	Num. of seeds	Learning rate	Weight decay	Scheduler	Batch size	Architecture	Optimizer	Domains in minibatch	Group by	Training epochs
CAMELYON17	10	1e-4	0	n/a	32	DenseNet121	SGD	3	Hospitals	10
FMOW	3	1e-4	0	n/a	32	DenseNet121	Adam	5	Times × regions	12

sive details on the WILDS datasets can be found in the corresponding WILDS paper (Koh et al., 2021).

The learned features are evaluated with V-REX and GroupDRO Sagawa et al. (2019), two representative SOTA OOD objectives in WILDS. In addition to OOD objectives, we evaluate the learned features with Deep Feature Reweighting (DFR) Kirichenko et al. (2022). DFR uses an additional OOD validation set where the spurious correlation does not hold to perform logistic regression based on the learned features. Intuitively, DFR can serve as a proper measure for the quality of learned invariant features Izmailov et al. (2022).

To ensure a fair comparison, our empirical approach strictly adheres to the experimental settings used by Chen et al. (2023a) in their analysis of the WILDS datasets listed in Table 6 and Table 7 and report the results.

We further use Integrated Gradients (Sundararajan et al., 2017) to compute attributions for each input feature with respect to the prediction of models trained by different algorithms. Integrated Gradients helps enhance the interpretability of complex models and understand why a model makes a certain prediction, which is as crucial as the prediction's accuracy, especially in sensitive and critical applications like healthcare, finance, and autonomous driving. By visualizing what the model is focusing on when making predictions, Integrated Gradients can help determine whether the model is considering the right features. The visualization is shown in Figure 8. The blue dots are the salient features. A deeper blue color denotes more salient features. It can be found that DOREEN is able to learn more meaningful and diverse features than ERM, BONSAI and FeAT.