Rich Feature Learning via Diversification

Supplementary Material

7. Proofs

We first introduce the notations used for theoretical analysis in Tab. 4

Table 4. 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. With slight abuse of notation, we also use Φ_i to refer to the i_{th} featurizer when the context allows.
$\omega = [\omega_1, \omega_2, ...\omega_n]^\top \in \mathbb{R}^n$	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

7.1. 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 $\underset{\omega}{\operatorname{argmin}} \mathcal{L}_{P_{tr}}(\omega, \phi) = \omega^*$, then for any real number λ:

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

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

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

□

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 $div(\Phi_1, \Phi_2) = \alpha^1_k * \sum_{i=1}^n \sum_{j=1}^n \mathbb{1}_{\Phi_{1i}=\Phi_{2j}}$ and intra-model part $div(\Phi_k) = \alpha^2_k * \sum_{1\leq i<j\leq n} \mathbb{1}_{\Phi_{ki}=\Phi_{kj}}$.

Lemma 7.1. *$div(\Phi_1, \Phi_2) = div(\Phi_1, \Phi_2/\Phi_{2t}) + div(\Phi_1, \Phi_{2t})$. 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 7.2. *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.

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

□

7.2. Proof of Proposition 1

(FeAT fails with a small μ) 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 $\ell(\omega^\top \Phi(x_i), y_i) \leq \theta$.*

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 (2) 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. □

7.3. Proof of Proposition 2

(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 2 .

Then for Φ_2, the loss by Equation (3) 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., $\alpha_2^1 > \delta - \lambda \geq 0$. □

7.4. 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 $\phi \notin \Phi$. Sec. 7.4 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 $\alpha_2^2 > n * \alpha_2^1$.

□

7.5. 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 $O(n_N^3)$.

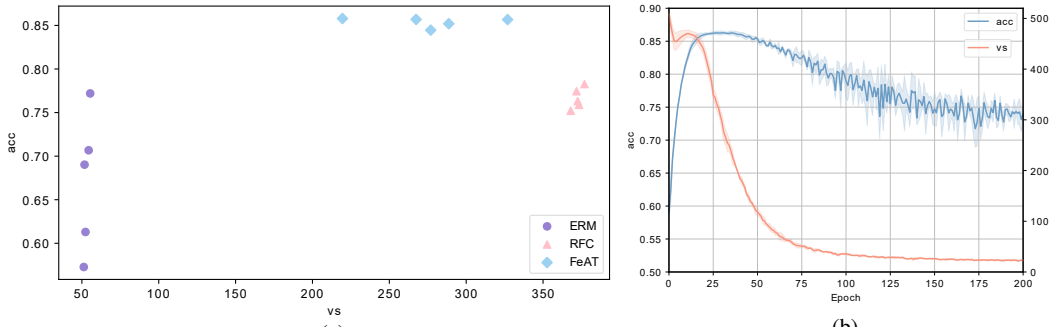
8. Related work

OOD generalization and Rich Feature Learning. There have been considerable efforts aimed at identifying the underlying causes of poor OOD generalization. Prior works [32, 41] highlight that spurious correlations—features informative on training data but not causally related to labels—significantly contribute to OOD failures. While RFL methods emphasize capturing a broad set of informative features (including potentially spurious ones) in the feature learning phase, as illustrated in Figure 1 (a), various OOD objectives designed to identify and filter out spurious features can be integrated into the complementary OOD training phase. Together, these two phases result in a model that captures a comprehensive feature set and correctly relies on invariant features for prediction. Recent studies [19, 53] further indicate that uncorrelated features may also degrade OOD performance, as coupling invariant and uncorrelated features can diminish the contributions of invariant features—a phenomenon referred to as feature contamination [53]. To address this, [19, 53] propose projecting the network's intermediate representations onto specific subspaces; however, reliably identifying these optimal subspaces remains challenging. Exploring the relationship between feature richness and feature contamination remains an open and promising direction for future research in the OOD generalization field.

Diversity-based approaches for OOD generalization. While not limited to OOD tasks, Yu et al. [50] promotes diverse and non-overlapping features in cascade classification by formulating feature selection as a submodular maximization problem to encourage the selection of complementary features that collectively enhance detection performance. Tschiatschek et al. [47] also proposes to learn mixtures of submodular functions to address the problem of image collection summarization. Wang et al. [48] introduces an attention diversity regularization for open-set recognition (OSR), encouraging mutually distinct attention maps across expert models to produce complementary representations and reduce open-space risk. Similarly, Pham and Plummer [35] enhances feature diversity in multi-channel imaging (MCI) tasks by employing a channel sampling strategy that promotes the selection of more distinct channel sets during training.

There has been a surge of efforts aimed at leveraging diversity to enhance OOD generalization. Several works focus on encouraging diversity at the output level. Teney et al. [46] enforces diversity through gradient disagreement among multiple classifiers sharing a common feature extractor. DivDis [31] encourages diversity by training multiple heads in a multi-headed neural network to make distinct predictions on unlabeled OOD samples while remaining accurate on in-distribution data. D-BAT [34] adopts a sequential training scheme where the first model is trained using standard ERM and the second model is trained with both ERM loss and an additional loss term to create disagreement with the first model's predictions on unlabelled samples from the OOD distribution. Both DivDis and D-BAT rely heavily on access to unlabeled OOD samples—a requirement that may not be practical in real-world settings. Although Teney et al. [46] avoids using OOD data during training, its model selection strategy still depends on such data to choose a classifier from the ensemble. To mitigate this limitation, Scimeca et al. [42, 43] propose using Diffusion Probabilistic Models (DPMs) to generate high-quality synthetic counterfactuals, serving as an alternative to OOD data. Rubinstein et al. [40] rather proposes a dynamic selection strategy for sampling OOD-like samples within the in-distribution dataset and further accelerate the training by stochastically selecting some pairs of ensemble models for disagreement in the optimization iterations. In contrast, DOREEN requires only the in-distribution training data, making it more broadly applicable and practical for OOD generalization. Moreover, Benoit et al. [8] critically examines existing diversity-based methods (including D-BAT and DivDis) and indicates that diversification alone is insufficient for effective OOD generalization, as many of the learned features may be spurious while these diversity-based methods lack mechanisms to identify and avoid relying on them. This limitation highlights the necessity of RFL methods like DOREEN, which involve the OOD training phase that select and rely on invariant features, thereby ensuring more robust OOD generalization.

Compared to existing diversity-based methods, DOREEN is not tied to any specific diversity measure as a flexible and principled framework that bridges diversity with rich feature learning. By promoting diversity within the feature space at



(a) (b)

Figure 5. 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.

both inter-model and intra-model levels, DOREEN enables the model to learn a broader spectrum of features to improve OOD generalization.

9. Limitations and Future Directions

In our experiments, DOREEN employs Determinantal Point Processes (DPP) to promote feature diversity. However, a practical limitation arises from potential numerical instability during the eigen-decomposition of the kernel matrix L when the bandwidth hyperparameter σ is set too large. Specifically, σ governs the Gaussian kernel used in DPP, which measures similarity between item pairs as $\exp\left(-\frac{|x_i - x_j|^2}{2\sigma^2}\right)$. When σ becomes excessively large, this similarity approaches 1 for all pairs, collapsing L into a near rank-deficient matrix. As a result, operations such as matrix inversion during eigen-decomposition can become numerically unstable. Nonetheless, DOREEN is designed as a flexible framework that bridges diversity and rich feature learning, without being tied to any specific diversity optimization method. This flexibility allows for the seamless integration of alternative diversity metrics, such as squared distance or cosine dissimilarity, to mitigate these numerical challenges while preserving the benefits of diverse feature learning.

Significant efforts have been devoted to understanding the causes of poor OOD generalization. Among these, spurious correlations—features that are predictive on the training distribution but not causally linked to the target labels—are widely recognized as a major contributor to OOD failures [32, 41]. RFL methods focuses primarily on spurious correlations as the source of OOD degradation, addressing this issue by encouraging the learning of a broad set of informative features (including potentially spurious ones) during the feature learning phase and various objectives can be employed to identify and filter out spurious features in the subsequent OOD training phase, as illustrated in Figure 1 (a). This two-stage approach enables models to capture a rich set of features while ultimately relying on invariant features for robust prediction. However, Idnani et al. [19], Zhang et al. [53] have shown that uncorrelated features, though not spurious, can also impair OOD performance, as coupling invariant and uncorrelated features can diminish the contributions of invariant features—a phenomenon referred to as feature contamination [53]. Exploring the relationship between feature richness and feature contamination remains an open and promising direction for future research in the OOD generalization field.

10. Detailed Experiments

In this section, we provide more details and the implementation, evaluation and hyperparameter setups in complementary to the experiments in Sec. 3 and Sec. 5. We conducted all the experiments utilizing NVIDIA GeForce RTX 3090.

10.1. More details about the experiments for motivating studies

Datasets. We conducted experiments on the ColoredMNIST dataset [4] with two environments for training and three environments for evaluation, 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) to generate the two training environments. That is, the target label

Table 5. 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	-

correlates with the invariant feature (the digit shape) with a probability 0.75. The spurious feature (color) correlates with the target label with a probability 0.9/0.8. The distinction between the two versions of the ColoredMNIST dataset lies in the feature-label correlation in the training environments: spurious (ColoredMNIST-025) or invariant(ColoredMNIST-01) features are better correlated with labels. However, in the OOD environments for testing, the spurious feature (color) correlates with the target label with a probability 0.9/0.5/0.1, and we report the worst accuracy of the model among these 3 environments. This testing protocol hits algorithms that rely on the spurious color feature because it happens to be more predictive than the robust feature in both training environments.

Architecture and optimization. We use a 4-layer MLP with a hidden dimension of 390/512 as the backbone model for all methods for the comparison of the ERM and two RFL methods/ERM training dynamics, where we take the first 3 layers as the featurizer and the last layer as the classifier, following the common practice [16, 25]. For experiments in Sec. 5.1 that assess the training dynamics of DOREEN, we train two models, each with a dimension of 256, and then concatenate them. For the optimization of the models, we use the Adam [23] optimizer with a learning rate of $1e-3$ and a weight decay of $1e-3$. For the featurizer trained by ERM, BONSAI and FeAT, we use V-REX [27] 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 [52], FeAT [10], our implementation follows [10]. For experiments in Sec. 3, we use V-REX [27] as the OOD objective to apply the OOD regularization and adopt the implementations from [52].

10.2. More details about the experiments in the controlled study

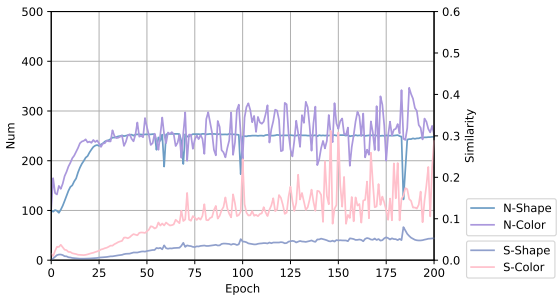
In this section we show the detailed empirical settings and more results in the controlled study on ColoredMNIST.

10.2.1. Details about the experimental settings

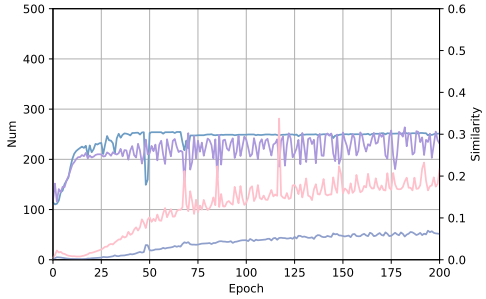
As for dataset, in addition to the original version where $\varepsilon_{tr} = \{(0.25, 0.1), (0.25, 0.2)\}$(denoted as as ColoredMNIST-025) and a modified ColoredMNIST (denoted as ColoredMNIST-01) with $\varepsilon_{tr} = \{(0.1, 0.2), (0.1, 0.25)\}$, we further utilize ColoredMNIST-sp, characterized 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 Sec. 10.1, except that we use a hidden dimension of 256 as in [10] to obtain a fair comparison.We report the mean and standard deviation of the performances of different methods with each configuration of hyperparameters 10 times with the random seeds from 1 to 10.

Implementation of feature learning and OOD training methods. For the common feature learning protocol with ERM and two SOTA RFL methods: BONSAI [52], FeAT [10], our implementation strictly follows [10]. For OOD objectives, we adopt the implementations from [52] for IRMv1 [4],V-REX [27] and IB-IRM [2]; and the implementations from [10] for IRMX [9].

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 [52] for BONSAI and [10] for FeAT. We train the model with BONSAI by 2 rounds with 50 epochs for round 1, 500 epochs for round 2, and 500 epochs for the synthesize round in ColoredMNIST-025. While in ColoredMNIST-01, round 1 contains 150 epochs, round 2 contains 400 epochs and the synthesize round contains 500 epochs. In ColoredMNIST-sp, BONSAI encounters issues due to an empty augmentation set. For the implementation of FeAT, we train the model with 2 rounds of FeAT in ColoredMNIST-025, 3 rounds of FeAT in ColoredMNIST-01, and 2 rounds of FeAT in in ColoredMNIST-sp, where each round contains 150 epochs. For the retain penalty, we follow Chen et al. [10] to use a fixed number of 0.01. ERM only contains 1 round, for which we train the model with 150 epochs in ColoredMNIST-025 as more epochs will incur severe performance degeneration [10]. While in ColoredMNIST-01, we train the model with ERM by 500 epochs



(a) ColoredMNIST-025



(b) ColoredMNIST-01

Figure 6. 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.

to match up the overall training epochs of FeAT and BONSAI. We provide a detailed distribution of the number of epochs in each round in Tab. 5. For the experiments on ColoredMNIST-025 and ColoredMNIST-01, we reported findings from Chen et al. [10] 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. Deep Ensemble, DIWA, and DOREEN utilize multiple models, while WLD-Reg promotes diversity within a single feature extractor, and Teney et al. [46] encourages diversity through gradient disagreement among multiple classifiers. For a fair comparison, using a 4-layer MLP as the backbone model where the first 3 layers / the last layer as the featurizer / classifier, we set Deep Ensemble, DIWA, and DOREEN to use two models (each with a hidden dimension of 256), and WLD-Reg to use one model with a hidden dimension of 512. For Teney et al. [46], we performed a grid search over 8, 16, 32 classifiers. WLD-Reg adopts the DPP loss as its diversity metric to align with DOREEN. All other experimental settings—such as training epochs and learning rates—were kept consistent across methods, with V-REX used as the common OOD training objective.

10.2.2. More empirical results

Comparison between ERM and RFL methods together with ERM training dynamics on ColoredMNIST-01. Figure 5(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 5(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 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[15]. The results, detailed in Figure 6, 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 5(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. [1], 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 7 further demonstrate that the OOD performance of a model is

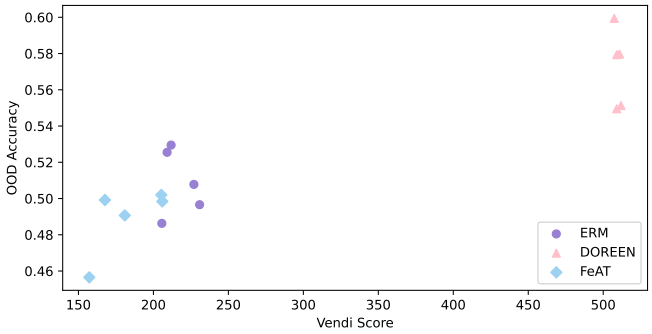


Figure 7. The feature diversity (x-axis) and OOD performance (y-axis) of featurizers trained with ERM, FeAT and DOREEN over five different random seeds. Post featurizer training, we measure the feature diversity using the Vendi Score, freeze the featurizer, and subsequently train a classifier using V-REX for OOD performance assessment.

Table 6. OOD performance of different diversity techniques.

	Deep Ensemble	DiWA	WLD-Reg	[46]	[46]*	DOREEN
ColoredMNIST-025	65.41 (±1.52)	59.01 (±3.97)	68.79 (±1.42)	65.30 (±3.09)	11.49 (±0.87)	**69.61** (±0.75)
ColoredMNIST-01	82.47 (±1.05)	70.47 (±7.01)	85.06 (±0.92)	84.01 (±1.19)	73.19 (±1.31)	**85.80** (±0.55)
ColoredMNIST-sp	51.46 (±1.88)	52.56 (±0.66)	55.26 (±3.12)	51.90 (±0.44)	9.86 (±0.43)	**56.66** (±2.34)

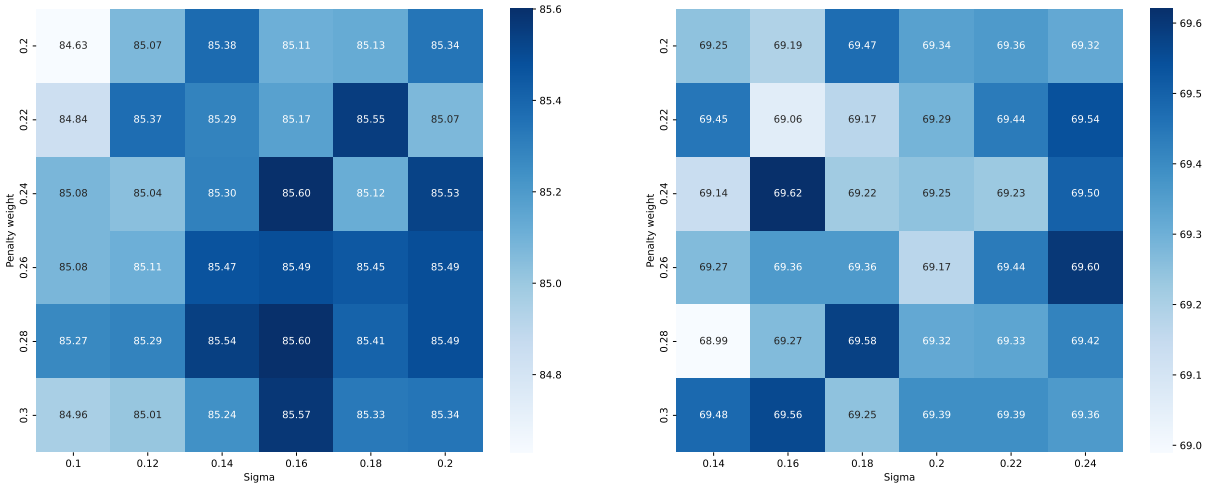
strongly correlated to the diversity of its featurizer. Moreover, in this extreme scenario, the feature diversity of FeAT-trained featurizer gains few improvement over that of ERM, leading to a almost the same or even worse OOD performance compared to ERM, while DOREEN effectively learns richer features and show apparently powerful performance.

Comparison with Other OOD Methods Incorporating Diversity Techniques. Benoit et al. [8] shows that diversity alone is not sufficient for OOD generalization—many learned features may be spurious, and no mechanism exists to filter them, while DOREEN addresses this by enabling the OOD training phase to identify and utilize invariant features as a RFL method.

We comprehensively cover diversity methods across all scopes—including featurizer, classifier, and whole-model level diversity—as illustrated in Figure 1(b). We additionally include a variant (denoted as [46]*) that applies their proposed classifier selection strategy, which uses OOD samples to select a classifier from the trained ensemble instead of retraining a new one for evaluation.

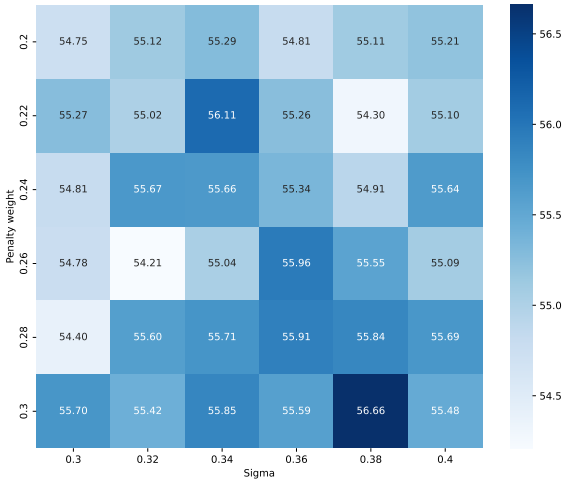
Even using data from exactly test distributions for selection, [46]* performs worse than [46] as proposed in the RFL paradigm. Consistent with Benoit et al. [8], [46] struggles to distinguish spurious features and avoid relying on them.

Hyper parameter tuning. In our experiments utilizing Determinantal Point Processes (DPP) to enforce diversity via a penalty term, two key hyperparameters require careful tuning: the diversity penalty weight α (introduced in Equation (5)) and the kernel bandwidth σ. Sigma serves as the hyperparameter for the DPP's kernel function, specifically a Gaussian kernel in our case, which governs how similarity (diversity) between item pairs is measured. For two elements x and y, the kernel similarity is defined as $\exp\left(-\frac{\|x_i - x_j\|^2}{2\sigma^2}\right)$. Intuitively, when examining a specific group of features, a smaller sigma value typically indicates a higher degree of diversity among them. In contrast, a larger sigma value suggests that the items are less diverse and encourages the model to push features farther apart. However, excessively large σ values risk collapsing the kernel matrix into a low-rank structure, which destabilizes DPP optimization by making the determinant computation numerically unstable. The key is to find a sigma that captures meaningful diversity. To determine a suitable σ, we first scaled it such that the resulting diversity loss had a magnitude comparable to the empirical risk minimization (ERM) loss, and then conducted a local grid search around this anchor to identify a well-performing setting that preserves both numerical stability and effective diversity. The penalty weight, on the other hand, strikes a balance between diversity and informativeness. Setting α too high risks overemphasizing diversity at the expense of selecting uninformative features, while setting it too low may fail to suppress redundant or spurious correlations. In our experiments, we fixed a predefined grid for the penalty weight α. Moreover, on ColoredMNIST-sp, where color strongly correlates with labels corresponding to a negligible λ, we further expanded the grid to include larger α values, motivated by insights from Proposition 2. The out-of-distribution

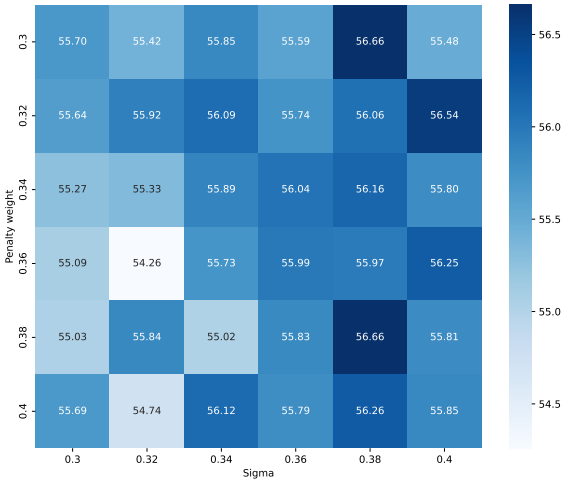


(a) ColoredMNIST-01

(b) ColoredMNIST-025



(c) ColoredMNIST-sp



(d) ColoredMNIST-sp

Figure 8. 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.

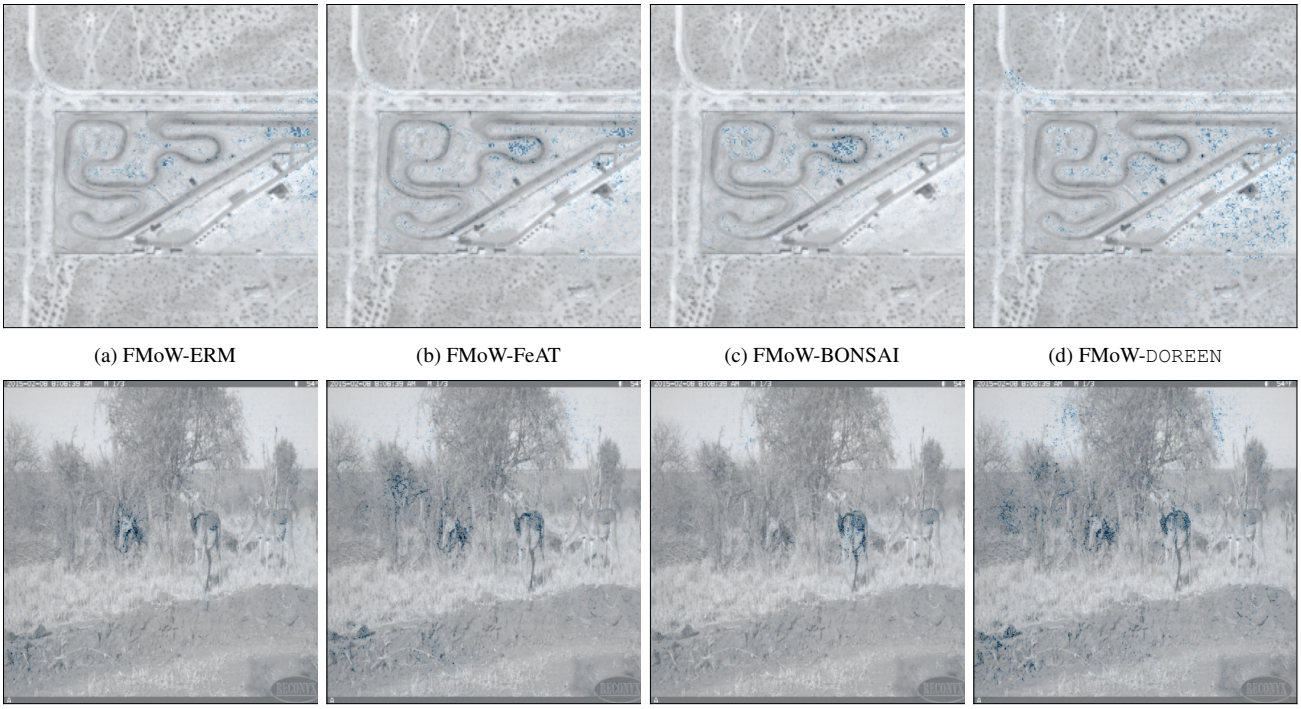
Table 7. Hyperparameter setups of feature learning algorithms for the experiments on WILDS

Dataset	Overall steps	Approx. epochs	Num. of rounds	Num. of models	Steps per round	Penalty weight	Sigma
CAMELYON17	10000	10	1	2	10000	0.25	0.15
FMOW	7000	4	1	2	7000	0.1	1.0
iWildCam	48000	10	1	2	48000	0.15	0.1

(OOD) performance, corresponding to various combinations of sigma and penalty weight, is illustrated in Figure 8. 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.

10.3. 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 three



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

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

Figure 9. Integrated Gradients visualization of models trained by ERM, BONSAI, FeAT and DOREEN on FMoW and iWildCam. The blue dots are the salient features. A deeper blue color denotes more salient features.

particularly challenging datasets: Camelyon17 [5], FMoW [12] and iWildCam [7] from the WILDS benchmark [25].

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. We will use the average accuracy as the evaluation metric and a DenseNet-121 [18] as the backbone for the featurizer. 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. The evaluation metric for FMoW is the worst region accuracy and the backbone model for the featurizer is a DenseNet-121 [18]. iWildCam is consist of 203,029 heat or motion-activated photos of animal specifies from 323 different camera traps across different countries around the world. The task of iWildCam is to classify the corresponding animal specifies in the photos. The domains is split according to the locations of the camera traps which could introduce the distribution shifts. We will use the Macro F1 as the evaluation metric and a ResNet-50 [17] as the backbone for the featurizer. Comprehensive details on the WILDS datasets can be found in the corresponding WILDS paper [25].

Table 8. 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	5
FMOW	3	1e-4	0	n/a	32	DenseNet121	Adam	5	Times $\times$ regions	3
iWildCam	3	1e-4	0	n/a	16	Resnet50	Adam	10	Trap locations	3

To ensure a fair comparison, our empirical approach strictly adheres to the experimental settings used by Chen et al. [10]

in their analysis of the WILDS datasets listed in Tab. 7 and Tab. 8 and report the results. We further use Integrated Gradients [45] 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 9. The blue dots are the salient features – more blue dots denotes more salient features. It can be found that DOREEN is able to learn more meaningful and diverse features than ERM, BONSAI and FeAT.