Ensemble Pruning for Out-of-distribution Generalization

Fengchun Qiao [1] **Xi Peng** [1]

Abstract

Ensemble of deep neural networks has achieved great success in hedging against single-model failure under distribution shift. However, existing techniques suffer from producing redundant models, limiting predictive diversity and yielding compromised generalization performance. Existing ensemble pruning methods can only guarantee predictive diversity for in-distribution data, which may not transfer well to out-of-distribution (OoD) data. To address this gap, we propose a principled optimization framework for ensemble pruning under distribution shifts. Since the annotations of test data are not available, we explore relationships between prediction distributions of the models, encapsulated in a topology graph. By incorporating this topology into a combinatorial optimization framework, complementary models with high predictive diversity are selected with theoretical guarantees. Our approach is model-agnostic and can be applied on top of a broad spectrum of off-the-shelf ensembling methods for improved generalization performance. Experiments on common benchmarks demonstrate the superiority of our approach in both multi- and single-source OoD generalization. The source codes are publicly available at: https://github.com/joffery/TEP.

1. Introduction

Despite the documented successes, the complex prediction rules learned by modern machine learning (ML) models, such as deep neural networks, are vulnerable to out-of-distribution (OoD) data. This means that an ML model, while highly accurate on average, may fail dramatically when faced with rare or unseen data distributions (Sagawa et al., 2019; Qiao et al., 2020). Although numerous algorithms have been proposed to improve OoD generalization, recent studies (Wiles et al., 2021; Gulrajani & Lopez-Paz, 2020; Koh et al., 2021) indicate that no single model consistently outperforms others across all theoretical or empirical contexts, and many perform worse than the standard baseline of empirical risk minimization (Vapnik, 1999).

Ensemble learning (Dong et al., 2020) emerges as a promising approach by leveraging the diversity of multiple models to mitigate the risk of single-model failure under distribution shifts. This diversity originates from variations in initialization (Lakshminarayanan et al., 2017), architectures (Li et al., 2022), and training processes (Wortsman et al., 2022; Ramé et al., 2023; Lin et al., 2024). However, due to the lack of access to data from target distributions during training, current methods tend to generate an excessive number of models in the hope of ensuring diversity at test time. These methods often lead to the creation of redundant models, and merely combining all models can reduce their complementary strengths against the target distribution, resulting in suboptimal generalization performance (see Fig. 1).

This underscores the importance of ensemble pruning (Tsoumakas et al., 2009), a process that involves selectively choosing an optimal subset from a pool of pre-trained models. Ensemble pruning aims to retain models that are most complementary to each other to enhance overall performance. Both theoretical analyses and empirical studies (Martinez-Munoz et al., 2008; Caruana et al., 2004) support the effectiveness of ensemble pruning in boosting the generalization ability of ensembles, a concept sometimes referred to as the "many-could-be-better-than-all" theorem (Zhou et al., 2002). Existing pruning methods typically utilize supervised diversity metrics based on training or validation data to inform the selection process. However, this approach encounters difficulties with data exhibiting distribution shifts, as the diversity metrics applicable to in-distribution data may not be relevant for out-of-distribution data. Consequently, the problem of efficiently pruning redundant models to improve performance when faced with distribution shifts remains an unexplored problem.

In this paper, we introduce a novel post-hoc optimization approach designed to prune redundant models at test time, thereby enhancing generalization to out-of-distribution

[1]DeepREAL Lab, Department of Computer and Information Sciences, University of Delaware, DE, USA. Correspondence to: Xi Peng <xipeng@udel.edu>.

Proceedings of the 41st International Conference on Machine Learning, Vienna, Austria. PMLR 235, 2024.

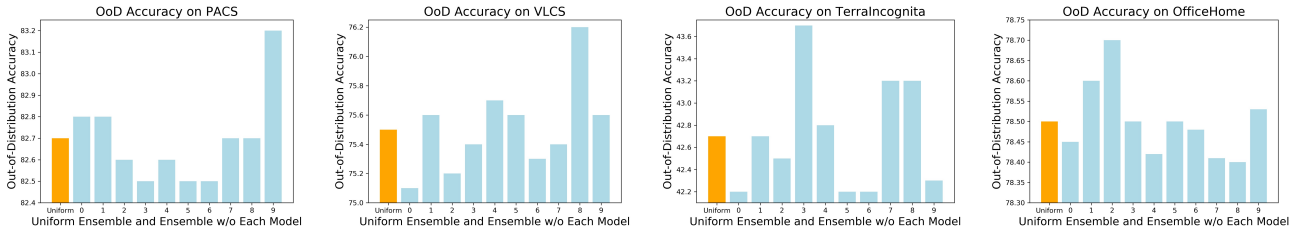

Figure 1. We evaluate the Out-of-Distribution (OoD) accuracy of a uniform ensemble and its variations using the *DomainBed* benchmark (Gulrajani & Lopez-Paz, 2020). To construct our ensemble, we employ the state-of-the-art *DiWA* (Rame et al., 2022) method, training ten models with varied hyperparameters. Our empirical analysis reveals selectively removing a particular model from the ensemble pool can enhance OoD performance. This finding suggests that model redundancy is a widespread issue in ensemble learning for OoD generalization, indicating the feasibility of creating more compact ensembles nonetheless exhibit superior generalization capabilities.

(OoD) data. The primary challenge is the selection of a complementary subset of models without the accessibility of data annotations. To overcome this challenge, we propose to learn an ensemble topology graph that captures the relationships between model predictions. This graph is integrated into a combinatorial optimization framework, enabling the adaptive selection of a diverse ensemble of models tailored to the test distribution. A distinctive feature of our method is its reliance solely on test data predictions for model selection, without necessitating any updates to model parameters. This approach sets it apart from test-time adaptation (Sun et al., 2020) strategies that typically involve re-training models. Our method is designed to be model-agnostic, making it compatible with a wide array of existing ensembling techniques to boost generalization performance. Through rigorous testing on well-established benchmarks, our method has demonstrated its superiority in improving generalization across both multi-source and single-source OoD scenarios. Our contributions are threefold:

- We pioneer the exploration of ensemble pruning in the context of distribution shifts, presenting a topology-informed post-hoc optimization framework that selectively curates a highly diverse ensemble for the test distribution.
- Our approach is model-agnostic and can be applied on top of a broad spectrum of off-the-shelf ensembling methods for improved generalization performance.
- Experiments on common benchmarks demonstrate the effectiveness of our method in enhancing OoD generalization, setting new standards in both multi-source and single-source contexts.

2. Preliminaries

The goal of ensemble pruning is to search for a good subset of members that performs as well as, or better than, the original ensemble. Error analysis of continuous problems (Breiman, 2001) shows that the ensemble error can be represented by a linear combination of the individual accuracy terms and pairwise diversity terms. Motivated by it, (Zhang et al., 2006) formulated ensemble pruning as a quadratic integer programming problem and utilized this linear combination as the optimization objective.

To achieve it, firstly a matrix P is used to record the error of all the member classifiers on the validation set:

$$P_{ij} = \begin{cases} 0, & \text{if classifier } j \text{ correctly predicts } i \\ 1, & \text{otherwise} \end{cases} \tag{1}$$

Thus, $G = P^\top P$ is a matrix with the interesting properties that the diagonal entries G_{ii} represent the misclassification errors made by each classifier i on the validation data, while the off-diagonal entries G_{ij} correspond to the number of common errors made by classifier i and j. Normalization is then applied on G to make its elements on the same scale:

$$\tilde{G}_{ii} = \frac{G_{ii}}{M}, \ \tilde{G}_{ij,i\neq j} = \frac{1}{2}\left(\frac{G_{ij}}{G_{ii}} + \frac{G_{ij}}{G_{jj}}\right), \tag{2}$$

where M is the number of data points. Intuitively, smaller $\tilde{G}_{ii}$ corresponds to a more accurate member classifier, while smaller $\tilde{G}_{ij}$ implies a more different classifier pair. Therefore, it is straightforward that a small value of $\sum_{ij} \tilde{G}_{ij}$ implies a good ensemble. Specifically, given N pre-trained models, selecting a sub-ensemble of size K ($K \leq N$) with both accuracy and diversity can now be formulated as:

$$\begin{aligned} &\min_{z} \ z^T \tilde{G} z \\ &\text{s.t.} \ \sum_i z_i = K, \ z_i \in \{0, 1\}. \end{aligned} \tag{3}$$

The binary variable z_i serves as a 0/1 weight or an indicator: when $z_i = 1$, the i th classifier will be selected. The parameter K controls the size of the pruned sub-ensemble and needs to be specified beforehand.

Compared to existing heuristics that use simple greedy search as the optimization method, solving Eq. 3 can rigorously reduce the generalization error with theoretical guarantee when there are no distribution shifts. However, this

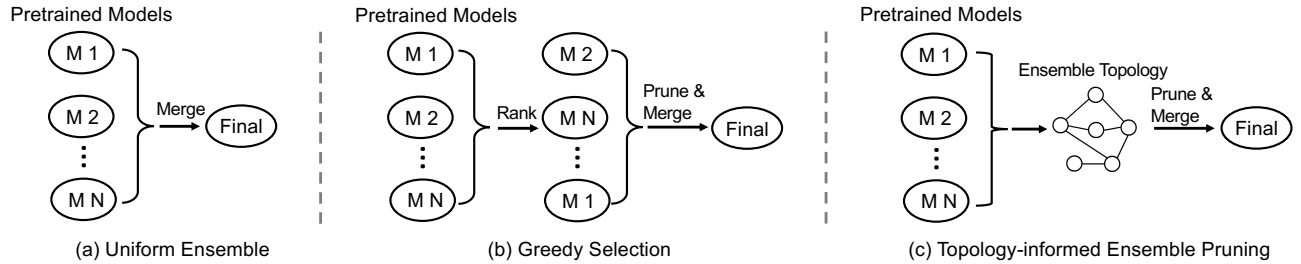

Figure 2. Overview of different strategies for ensemble pruning. Different from existing works that are tailored for in-distribution data, our proposed topology-informed ensemble pruning is capable of promoting predictive diversity and lowering generalization errors on shifted distributions with theoretical guarantees.

method as well as other alternative ensemble pruning methods suffer from critical limitations under distribution shifts: they typically utilize supervised diversity metrics based on training/validation data to inform the selection process; the diversity on in-distribution data may not be transferred to out-of-distribution data. Consequently, the problem of efficiently pruning redundant models to improve performance under distribution shifts remains an unexplored problem.

3. Topology-informed Ensemble Pruning

It is non-trivial to select a subset of complementary models for out-of-distribution (OoD) data as the annotations are not available. To bridge this gap, we propose a topology-informed optimization method for ensemble pruning under distribution shifts. Our key idea is to explore an ensemble topology that captures the relationships between model predictions. By incorporating this topological information into the optimization in Eq. 3, we can theoretically guarantee a selection of models with improved diversity on shifted test distributions, thereby improving OoD robustness. Our method consists of two steps: *Topology Learning* (Sec. 3.1) and *Learning on Topology* (Sec. 3.2).

3.1. Topology Learning

We represent the ensemble topology as a weighted graph $\mathcal{G} = (V, E, W)$ that captures predictive relationships between models. The nodes V symbolize individual models, the edges E depict connectivities between models, and the adjacency matrix W contains edge weights. The edge weight between models i and j is defined as: $W_{ij} = \exp(-D_{ij}^2/2)$ where D_{ij} represents the distance between models i and j. We employ the discrepancy between Fisher Information Matrices (FIM) (Fisher, 1922) as the distance metric due to its sensitivity in capturing parameter-output relationships. This allows us to assess model similarities in their prediction mechanisms, laying the foundation for further analysis of model redundancy.

For a network with parameters θ, the Fisher matrix F_θ is a positive semi-definite matrix that expresses how changes to θ impact the output distribution:

$$F_\theta = \mathbb{E}_x[\mathbb{E}_{y\sim p_\theta(y|x)}[\nabla_\theta \log p_\theta(y|x)(\nabla_\theta \log p_\theta(y|x))^T]]. \tag{4}$$

The full Fisher matrix is intractable to compute and store for large networks. A common approximation uses the diagonal Fisher (Matena & Raffel, 2022), which has a similar cost to backpropagating gradients over N examples. Specifically, we estimate the diagonal Fisher F_θ via:

$$\hat{F}_\theta = \frac{1}{M}\sum_{i=1}^{M} \mathop{\mathbb{E}}_{y\sim p_\theta(y|x_i)} \left(\nabla_\theta \log p_\theta\left(y \mid x_i\right)\right)^2. \tag{5}$$

In Eq. 5, the gradients are computed over M inputs sampled from the test set.

Alternative distance metrics. In our preliminary experiments, we have explored several common distance metrics including ℓ_2 distance, Earth Mover's Distance (EMD), and Maximum Mean Discrepancy (MMD). We empirically found these metrics are not as effective as FIM (see Fig. 6). This indicates the parameter-output sensitivity plays a key role in topology learning.

3.2. Learning on Topology

After acquiring the ensemble topology $\mathcal{G}$, we integrate $\mathcal{G}$ into the integer programming in Eq. 3 and formulate the optimization problem as:

$$\begin{aligned} \min_z \quad & \underbrace{z^T\tilde{G}z}_{\text{In-Distribution}} + \lambda \underbrace{z^T W z}_{\text{Out-of-Distribution}} \\ \text{s.t.} \quad & \sum_i z_i = K,\ z_i \in \{0,1\}. \end{aligned} \tag{6}$$

Here, λ is a hyperparameter to balance in-distribution accuracy and out-of-distribution diversity. Direct solution of the problem in Eq. 6 is challenging due to its NP-hard nature and binary constraints. To make it tractable, we propose to relax the binary constraints by replacing $z_i \in \{0, 1\}$ with $0 \le z_i \le 1$. This changes the problem from a combinatorial optimization problem to a continuous one, which is

easier to solve. To further simplify, we employ semi-definite relaxation by introducing a matrix $Z = zz^T$:

$$\begin{aligned} &\min_{Z} \ \mathrm{tr}(\tilde{G}Z) + \lambda\,\mathrm{tr}(WZ) \\ &\text{s.t. } \ \mathrm{tr}(Z) = K,\ Z \succeq zz^T,\ Z \succeq 0. \end{aligned} \tag{7}$$

Eq. 7 can be solved by standard semi-definite programming (SDP) solvers. After solving the relaxed continuous optimization problem, we obtain a relaxed solution matrix Z. To convert Z into a discrete 0-1 solution z that selects the models, we apply an efficient rounding procedure. Specifically, we pick the K indices corresponding to the K largest values in the diagonal of Z and set them to 1 in z, while setting all other entries to 0. We empirically found the SDP-relaxation yields almost the same OoD accuracy as brute force (see Tab. 5), indicating the relaxation provides a tight approximation to the original combinatorial problem.

The proposed topology-informed optimization framework seamlessly incorporates ensemble topology and accuracy information to select a subset of models suited for OoD data. By relaxing the integer constraints, we transform the intractable problem into an efficiently solvable SDP. The obtained continuous solution is rounded to retrieve a 0-1 model selection vector z. Our approach provides the following key advantages: (i) Jointly optimizes for OoD diversity and in-distribution accuracy without separate steps. (ii) Agnostic to the model family and applicable to any pre-trained models. (iii) Efficient test-time computation that only requires forwarding samples through models without re-training. In summary, the ensemble topology graph captures predictive relationships, while the Gram matrix provides accuracy validation. Fusing both sources of information into a combinatorial optimization formulation allows intelligently pruning the ensemble to improve robustness under distribution shift. In Sec. 4, we show that the pruned ensemble obtained by solving the quadratic programming problem in Eq. 7 leads to promoted diversity (Lem. 4.1) and lower generalization error (Thm. 4.4) on target distributions.

4. Theoretical Analysis

In this section, we provide theoretical guarantees for the proposed topology-aware ensemble pruning. In Lem. 4.1 (Diversity Promotion), we show the pruned ensemble promotes diversity compared to the original set of models. In Thm. 4.4 (Generalization Error of the Pruned Ensemble), we present a bias-variance-diversity decomposition of the expected ensemble risk, which highlights the role of diversity in reducing the generalization error on target distributions. We provide the step-by-step proof in the Appendix.

Lemma 4.1 (Diversity Promotion)**.** *Let $S \subseteq V$ be the pruned ensemble with the size of K, obtained by solving the optimization problem in Eq. 6. Define the average pairwise similarity among all models in V as:*

$$W_V = \frac{2}{N(N-1)} \sum_{i<j} W_{ij}, \tag{8}$$

and the average pairwise similarity among models in the pruned ensemble S as:

$$W_S = \frac{2}{K(K-1)} \sum_{i<j} W_{ij} z_i z_j. \tag{9}$$

Then, the following statement holds:

$$1 - W_S \geq 1 - W_V. \tag{10}$$

Remark. Lem. 4.1 shows that the average pairwise distance among models in the pruned ensemble S is higher than the average pairwise distance among all models in V. This implies that the pruned ensemble promotes diversity compared to the original set of models.

Definition 4.2 (Bias-Variance Decomposition. Adapted from (Geman et al., 1992))**.** Given a loss function ℓ, we define $\tilde{Y} \overset{\text{def}}{=} \arg\min_{Y \in \mathcal{Y}} \mathbb{E}_D[\ell(\hat{Y}, Y)]$ as the centroid of the model distribution. The bias-variance decomposition is formulated as:

$$\begin{aligned} &\underbrace{\mathbb{E}_D[\mathbb{E}_{X,Y}[\ell(\hat{Y}, Y)]]}_{\text{Expected risk}} = \mathbb{E}_X[\underbrace{\mathbb{E}_{Y|X}[\ell(Y^*, Y)]}_{\text{Noise}} \\ &+ \underbrace{\ell(\tilde{Y}, Y^*)}_{\text{Bias}} + \underbrace{\mathbb{E}_D[\ell(\hat{Y}, \tilde{Y})]}_{\text{Variance}}], \end{aligned} \tag{11}$$

where the conditional mean $Y^* \overset{\text{def}}{=} \mathbb{E}_{Y|X}[Y]$ is the Bayes-optimal prediction.

Definition 4.3 (Centroid Combiner Rule)**.** For a test point (x, y), given a set of model predictions $\{\hat{Y}_i\}_{i=1}^K$, the centroid combiner $\overline{Y}$ is the minimizer of the averaged loss $\ell(\hat{Y}_i, Y)$, over all ensemble members:

$$\overline{Y} \overset{\text{def}}{=} \arg\min_{Y \in \mathcal{Y}} \frac{1}{K} \sum_{i=1}^{K} \ell(\hat{Y}_i, Y). \tag{12}$$

Theorem 4.4 (Generalization Error of the Pruned Ensemble. Adapted from (Wood et al., 2023))**.** *Given a set of model predictions $\{\hat{Y}_i\}_{i=1}^K$ and a loss function ℓ, assuming a bias-variance decomposition holds in Def. 4.2, the following decomposition also holds:*

$$\begin{aligned} &\underbrace{\mathbb{E}_D[\mathbb{E}_{X,Y}[\ell(\overline{Y}, Y)]]}_{\textit{Expected ensemble risk}} = \mathbb{E}_X[\underbrace{\mathbb{E}_{Y|X}[\ell(Y^*, Y)]}_{\textit{Noise}} + \underbrace{\frac{1}{K}\sum_{i=1}^{K} \ell(\tilde{Y}_i, Y^*)}_{\textit{Average bias}} \\ &+ \underbrace{\frac{1}{K}\sum_{i=1}^{K} \mathbb{E}_D[\ell(\hat{Y}_i, \tilde{Y}_i)]}_{\textit{Average variance}} - \underbrace{\mathbb{E}_D[\frac{1}{K}\sum_{i=1}^{K} \ell(\hat{Y}_i, \overline{Y})]}_{\textit{Diversity}}], \end{aligned} \tag{13}$$

Method	VLCS					Terra Incognita				
	Caltech101	LabelMe	SUN09	VOC2007	Avg.	Loc.100	Loc.38	Loc.43	Loc.46	Avg.
ERM	97.7 ± 0.4	64.3 ± 0.9	73.4 ± 0.5	74.6 ± 1.3	77.5	49.8 ± 4.4	42.1 ± 1.4	56.9 ± 1.8	35.7 ± 3.9	46.1
SagNet	97.9 ± 0.4	64.5 ± 0.5	71.4 ± 1.3	77.5 ± 0.5	77.8	53.0 ± 2.9	43.0 ± 2.5	57.9 ± 0.6	40.4 ± 1.3	48.6
RSC	97.9 ± 0.1	62.5 ± 0.7	72.3 ± 1.2	75.6 ± 0.8	77.1	50.2 ± 2.2	39.2 ± 1.4	56.3 ± 1.4	40.8 ± 0.6	46.6
SWAD	98.8 ± 0.1	63.3 ± 0.3	75.3 ± 0.5	79.2 ± 0.6	79.1	55.4 ± 0.0	44.9 ± 1.1	59.7 ± 0.4	39.9 ± 0.2	50.0
DENS (Lakshminarayanan et al., 2017) (Average Model Outputs)										
+ Uniform	98.7 ± 0.5	64.2 ± 0.4	73.4 ± 0.3	78.6 ± 1.2	78.7	53.3 ± 1.6	42.8 ± 1.1	60.5 ± 0.8	41.0 ± 0.5	49.4
+ Greedy	98.8 ± 0.1	64.3 ± 0.4	72.6 ± 0.5	79.7 ± 0.7	78.9	52.8 ± 1.0	43.1 ± 0.7	59.2 ± 0.3	40.2 ± 0.7	48.8
+ Ours	98.8 ± 0.2	64.6 ± 0.5	74.9 ± 0.1	79.5 ± 0.7	**79.5**	55.6 ± 1.6	43.9 ± 0.8	60.0 ± 0.7	42.5 ± 0.3	**50.5**
DiWA (Rame et al., 2022) (Average Model Weights)										
+ Uniform	98.7 ± 0.2	64.1 ± 0.3	75.5 ± 0.1	79.1 ± 0.2	79.3	55.3 ± 2.0	45.3 ± 0.2	62.0 ± 0.4	42.7 ± 0.3	51.3
+ Greedy	98.9 ± 0.1	64.5 ± 0.2	72.8 ± 0.1	80.6 ± 0.5	79.2	54.4 ± 1.3	46.0 ± 0.3	60.1 ± 0.3	41.0 ± 0.5	50.4
+ Ours	99.0 ± 0.1	64.6 ± 0.3	76.5 ± 0.3	79.6 ± 0.2	**79.9**	57.0 ± 1.6	46.7 ± 0.2	60.7 ± 0.7	43.9 ± 0.5	**52.1**
Model Ratatouille (Ramé et al., 2023) (Average Model Weights)										
+ Uniform	99.3 ± 0.0	60.8 ± 0.3	74.3 ± 0.3	79.5 ± 0.3	78.5	57.9 ± 0.2	50.1 ± 0.7	59.8 ± 0.1	38.9 ± 0.5	**51.8**
+ Greedy	99.0 ± 0.0	62.4 ± 0.5	73.8 ± 0.3	79.5 ± 0.1	78.7	54.0 ± 2.0	47.7 ± 0.8	57.3 ± 0.8	37.9 ± 1.2	49.2
+ Ours	99.0 ± 0.1	62.8 ± 0.3	76.1 ± 0.4	79.6 ± 0.1	**79.4**	58.4 ± 0.9	49.6 ± 0.5	58.4 ± 0.1	40.3 ± 0.8	51.7

Table 1. Accuracy (%) of multi-source out-of-distribution generalization on *Terra Incognita* (Beery et al., 2018) and *VLCS* (Fang et al., 2013) datasets. The models are tested on the specified distribution and trained on all the other distributions. Our method consistently outperforms uniform ensemble and greedy selection on both datasets. We highlight the **best results** and the second best results. Results of other baselines are from Gulrajani and Lopez-Paz (Gulrajani & Lopez-Paz, 2020). The results are an average of three trials.

where $\tilde{Y}_i \overset{def}{=} \arg\min_{Y \in \mathcal{Y}} \mathbb{E}_D[\ell(\hat{Y}_i, Y)]$. *Eq. 18 decomposes the expected ensemble risk into the noise, the average bias, the average variance, and the diversity.*

Remark. Thm. 4.4 presents the bias-variance-diversity decomposition of the expected ensemble risk. It shows the diversity term, which measures the expected disagreement among ensemble members, reduces the overall risk. This highlights the importance of promoting diversity in ensemble pruning, as achieved by the proposed topology-aware method. However, maximizing OoD diversity alone might not suffice, as it could lead to the selection of models with lower OoD accuracy, thereby increasing the average bias. The challenge lies in choosing models with high OoD accuracy without access to target data during training or annotation during testing. Recent studies (Miller et al., 2021) indicate a strong correlation between out-of-distribution performance and in-distribution performance. Thus, we prioritize models with high in-distribution accuracy to potentially mitigate the average bias, which justifies the in-distribution accuracy term included in Eq. 6.

5. Experiments

We evaluate our method on the common OoD generalization benchmark *DomainBed* (Gulrajani & Lopez-Paz, 2020). We conduct experiments on both multi- and single-source (Peng et al., 2024) out-of-distribution generalization. Following (Gulrajani & Lopez-Paz, 2020), we use a validation set selected from the training domains for model selection and all the experimental results are averaged over 3 trials.

Baselines. We evaluate our method on top of three representative ensemble methods: DENS (Lakshminarayanan et al., 2017), Model Ratatouille (Ramé et al., 2023), and DiWA (Rame et al., 2022). In DENS, the models are trained with different initialization. In DiWA, the models are obtained from independent runs that differ in hyperparameters, data augmentations, and batch orders. To ensure the model weights are averageable, the models share the same pre-trained initialization and use a mild range of hyperparameters. In Model Ratatouille, the models are trained with diverse auxiliary tasks.

We compare our method with Uniform Ensemble, Greedy Selection, and traditional ensemble pruning methods. Greedy selection is a popular method for model selection in OoD generalization and has been adopted in (Rame et al., 2022; Ramé et al., 2023; Wortsman et al., 2022): models are ranked in decreasing order of validation accuracy and sequentially added only if they improve the ensemble's validation accuracy. This method is *a.k.a.* restricted selection. traditional ensemble pruning methods can be broadly categorized into three primary families: ranking-based (Guo & Boukir, 2013), clustering-based (Onan et al., 2017), and optimization-based (Bian et al., 2019). We compare with several representative methods: DREP (Li et al., 2012), DivP (Cavalcanti et al., 2016), Spectral (et al, 2014), SDP (Zhang et al., 2006), and COMEP (Bian et al., 2019).

We also compare with other single-model based OoD baselines, including ERM (Vapnik, 1999), SagNet (Nam et al., 2021), RSC (Huang et al., 2020), and SWAD (Cha et al., 2021). Following (Gulrajani & Lopez-Paz, 2020; Koh et al., 2021), we use a ResNet-50 (He et al., 2016) pre-trained on

		Order-based		Clustering-based		Optimization-based		
Dataset	Uniform	Greedy	DREP	DivP	Spectral	SDP	COMEP	Ours
Terra Incognita	51.3	50.4	50.6	50.8	49.8	50.9	<u>51.5</u>	**52.1**
VLCS	<u>79.3</u>	79.2	78.9	78.5	78.0	79.2	78.7	**79.9**

Table 2. Accuracy (%) of ensemble pruning methods on *Terra Incognita* (Beery et al., 2018) and *VLCS* (Fang et al., 2013) datasets. We found existing ensemble pruning methods cannot substantially outperform uniform ensemble under distribution shifts. This indicates that in-distribution diversity does not transfer well to out-of-distribution data.

	Terra Incognita				VLCS			
Method	Loc.100	Loc.38	Loc.46	Avg.	Caltech101	SUN09	VOC2007	Avg.
ERM	36.5	30.6	37.4	34.8	89.0	52.9	63.4	68.5
SagNet	31.0	39.0	39.1	36.4	88.5	55.6	65.8	70.0
RSC	38.3	31.4	39.4	36.4	88.3	55.6	64.5	69.4
SWAD	34.0	35.5	44.7	38.0	82.5	58.6	70.8	70.6
DiWA (Rame et al., 2022) (Average Model Weights)								
+ Uniform	38.2	35.0	44.4	<u>39.2</u>	89.7	58.8	70.2	<u>72.9</u>
+ Greedy	35.8	33.3	44.1	37.7	85.2	57.2	72.5	71.6
+ Ours	38.9	35.7	44.9	**39.8**	90.3	59.5	71.0	**73.6**

Table 3. Accuracy (%) of single-source out-of-distribution generalization on *Terra Incognita* (Beery et al., 2018) and *VLCS* (Fang et al., 2013) datasets. The models are trained on the specified distribution and tested on all the other distributions. In *Terra Incognita*, we train models on "Loc.43" and test on other locations; In *VLCS*, we train models on "LabelMe" and test on other datasets.

ImageNet (Russakovsky et al., 2015) as the backbone for all experiments.

5.1. Multi-source OoD Generalization

In multi-source OoD generalization, the models are tested on the specified distribution and trained on all the other distributions (domains). We show the detailed results on *VLCS* and *Terra Incognita*. *VLCS* contains photographic images from four domains: Caltech101, LabelMe, SUN09, and VOC2007. There are 10,729 total images with dimensions of (3, 224, 224) pixels across 5 classes. *Terra Incognita* consists of photos of wild animals captured by camera traps at four different locations. The dataset contains 24,788 images of size (3, 224, 224) pixels from 10 different classes. The results are shown in Tab. 1. On both datasets, DiWA with our approach achieves the best average test accuracy. On *VLCS* dataset, it outperforms uniform ensemble by 0.6% and greedy selection by 0.7%. More significantly, on the challenging *Terra Incognita* dataset, it surpasses uniform ensemble by 0.8% and greedy selection by 1.7%. The results demonstrate the effectiveness of our method for improving generalization across diverse domains. We compare with traditional ensemble pruning methods. Results on *Terra Incognita* and *VLCS* are shown in Tab. 2. We found existing ensemble pruning methods cannot substantially outperform uniform ensemble under distribution shifts. This indicates in-distribution diversity does not transfer well to out-of-distribution data.

5.2. Single-source OoD Generalization

In single-source OoD generalization (Qiao & Peng, 2021), the models are trained on the specified distribution and tested on all other distributions. Specifically, in *Terra Incognita*, we train models on "Loc.43" and test on other locations; In *VLCS*, we train models on "LabelMe" and test on other datasets. Since most OoD generalization methods requires multiple domains for training, in this experiments, we only compare our method with ERM, SagNet, RSC and SWAD. We show the detailed results on *VLCS* and *Terra Incognita* in Tab. 3. As observed, DiWA+Ours achieves the best average accuracy of 39.8% on *Terra Incognita* and 73.6% on *VLCS*, outperforming both uniform and greedy DiWA selection. Compared to prior domain generalization methods like ERM, SagNet, RSC and SWAD, DiWA+Ours delivers competitive or superior performance. The consistent gains over alternative selection strategies demonstrate the benefit of the proposed approach for selecting diverse ensembles that can generalize from a single training domain.

5.3. Ablation Study

In this section, we conduct ablation studies on the hyperparameter λ and the number of models to retain K. Additionally, we compare the performance between the SDP-relaxation with brute force (exhaustive search), and visualize the relation between prediction diversity vs. out-of-distribution accuracy.

	VLCS					Terra Incognita				
Method	Caltech101	LabelMe	SUN09	VOC2007	Avg.	Loc.100	Loc.38	Loc.43	Loc.46	Avg.
Knee Point	98.8	64.3	76.5	79.4	79.8	56.9	46.7	60.7	44.3	52.2
$\lambda = 1$	99.0	64.6	76.5	79.6	79.9	57.0	46.7	60.7	43.9	52.1

Table 4. Accuracy (%) of Pareto optimization and $\lambda = 1$. Results indicate that setting $\lambda = 1$ is a simple yet hard-to-beat baseline, despite the potential benefits of using Pareto optimization for automatic hyperparameter selection.

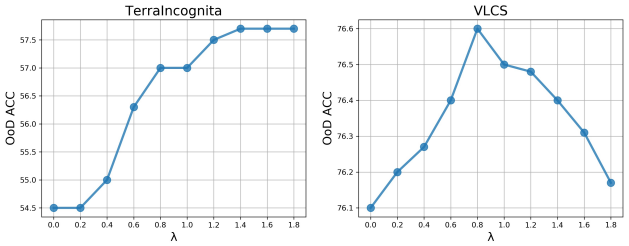

Figure 3. Ablation study on λ. In *Terra Incognita*, we found in-distribution accuracy does not contribute to the improvement. We suspect the reason is that, when distribution shift is large, the in-distribution accuracy may not be informative for OoD data. In *VLCS*, we found both in-distribution accuracy and OoD diversity contribute to the improvement.

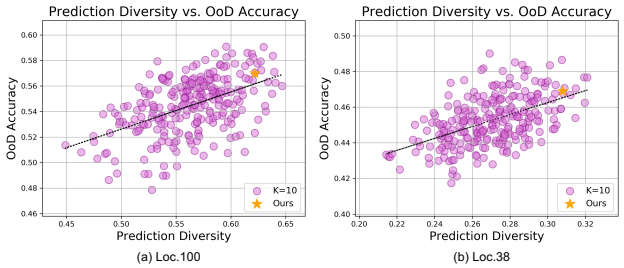

Figure 4. Visualization of prediction diversity vs. out-of-distribution accuracy on *Terra Incognita*. Each circle denotes a sub-ensemble with a size of 10 (half the size of the ensemble pool). The sub-ensemble selected by our method surpassed most of the selection combinations in terms of both prediction diversity and OoD accuracy.

Hyperparamter λ. λ is used to balance in-distribution accuracy and out-of-distribution diversity. We report the results under different λ on *Terra Incognita* and *VLCS*. The results are shown in Fig. 3. In *Terra Incognita*, we found large λ contributes to higher out-of-distribution performance. It indicates in-distribution accuracy does not contribute to the improvement. We suspect the reason is that, when distribution shift is large, the in-distribution accuracy/diversity may not be informative for out-of-distribution data. In *VLCS*, we found both in-distribution accuracy and out-of-distribution diversity contribute to the improvement. We empirically set $\lambda = 1$ for all experiments. To automatically choose λ, we explore Pareto optimization (Qian et al., 2015). Pareto optimization is a technique for solving problems with multiple conflicting objectives, which in our case are maximizing both the in-distribution accuracy and the out-of-distribution diversity. By applying Pareto optimization to Eq. 7, we aim to find a set of optimal solutions, called the Pareto set, that represents the best trade-offs between these objectives. Each solution in the Pareto set corresponds to a different value of λ, allowing us to explore and select suitable values without manual tuning. Following the approach in (Maltese et al., 2016), we choose the knee point, which is considered the most "cost-effective" point, from the Pareto set. To evaluate the effectiveness of this approach, we compare the performance of the knee point with the fixed value of $\lambda = 1$. The results on *Terra Incognita* and *VLCS* datasets are shown in Tab. 4. As observed, the knee point only outperforms $\lambda = 1$ on "Loc. 46" while being equal to or inferior to $\lambda = 1$ in all other distributions. These results indicate that setting $\lambda = 1$ is a simple yet hard-to-beat baseline, despite the potential benefits of using Pareto optimization for automatic hyperparameter selection.

Number of models to retain K. In Fig. 5, we show the OoD accuracy on *Terra Incognita* (Beery et al., 2018) with the different numbers of models to retain (K). In locations other than "Loc.43", the accuracy increases first and then decreases when we increase K. However, in "Loc.43", the best performance is achieved at $K = N = 20$. In this case, the results indicate that there is no redundancy in the ensemble pool and each model is necessary for the ensemble. Note that the accuracy on Loc.43 is the highest among all the locations. This indicates the distribution shift (between the source and target) is not as prominent as other locations. The results indicate our method yields better performance when the distribution shifts are relatively significant. Following (He et al., 2024), we set $K = \lfloor N/2 \rfloor$ for all experiments.

SDP-relaxation vs. Brute force. Since Eq. 6 is NP-hard, we quantitatively compare the performance between the SDP-relaxation (Eq. 7) with brute force (exhaustive search). The results on Terra and VLCS datasets are shown in Tab. 5. As seen, the SDP-relaxation yields almost the same OoD accuracy as brute force, indicating the relaxation provides a tight approximation to the original combinatorial problem.

Prediction diversity vs. OoD accuracy. We investigate the relation between prediction diversity vs. out-of-distribution accuracy on *Terra Incognita*. Given the ensemble pool with

Method	VLCS					Terra Incognita				
	Caltech101	LabelMe	SUN09	VOC2007	Avg.	Loc.100	Loc.38	Loc.43	Loc.46	Avg.
Brute Force	99.0	64.7	76.7	79.6	80.0	57.0	46.7	60.7	44.0	52.1
SDP-relaxation	99.0	64.6	76.5	79.6	79.9	57.0	46.7	60.7	43.9	52.1

Table 5. Accuracy (%) of SDP-relaxation and brute force. The SDP-relaxation yields almost the same OoD accuracy as brute force, indicating the relaxation provides a tight approximation to the original combinatorial problem.

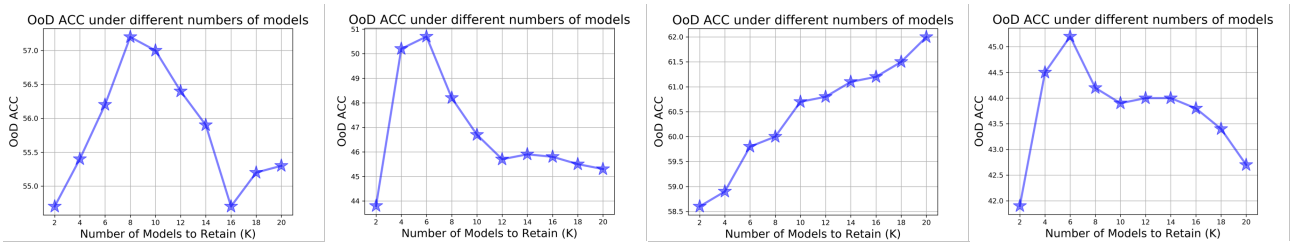

Figure 5. Ablation Study of K (Number of Models to Retain) on *Terra Incognita* Dataset. In locations other than "Loc.43", the accuracy increases first and then decreases when we increase the number of models to keep (K). However, in "Loc.43", the best performance is achieved at $K = N = 20$. In this case, the results indicate that there is no redundancy in the ensemble pool and each model is necessary for the ensemble. Note that the accuracy on Loc.43 is the highest among all the locations. This indicates the distribution shift (between the source and target) is not as prominent as other locations. The results indicate our method yields better performance when the distribution shifts are relatively significant.

the size of 20, we calculate the prediction diversity and out-of-distribution accuracy of each sub-ensemble with the size of 10 ($K = 10$). Following (Rame et al., 2022; Ramé et al., 2023), we use ratio-error (Aksela, 2003) as the diversity metric. It calculates the ratio of error diversity measured between a pair of classifiers. A higher value means that the base classifiers are less likely to make the same errors. Ratio-error is defined as $\frac{N^{01}+N^{10}}{N^{00}}$, where N^{ij} is the number of times that the first classifier is (correct if $i = 1$ or wrong if $i = 0$) and the second classifier is (correct if $j = 1$ or wrong if $j = 0$). We calculate the mean of all pairwise ratio-error within each sub-ensemble. The visualization is shown in Fig. 4. We empirically found that diversity and accuracy are linearly correlated. The sub-ensemble selected by our method outperforms the vast majority of the remaining ones, indicating the effectiveness of our method in encouraging predictive diversity for improved OoD performance.

6. Related Work

OoD Generalization and Ensemble learning. The goal of OoD generalization is to generalize the model from source distributions to unseen target distributions. Existing methods can be classified into two categories: invariant learning (Arjovsky et al., 2019; Yang et al., 2023; Li et al., 2023) and robust optimization (Sagawa et al., 2019; Qiao & Peng, 2023; Ma et al., 2024). However, recent studies (Wiles et al., 2021; Gulrajani & Lopez-Paz, 2020; Koh et al., 2021) have demonstrated that no single model can consistently achieve superior performance across all OoD scenarios. A series of methods based on ensemble learning (Pagliardini et al., 2023; Lee et al., 2023; Rame et al., 2022; Wortsman et al., 2022) have been proposed to consistently improve the OoD performance. Diversity plays a key role in ensemble learning as the error decreases with the covariance of ensemble members (Ueda & Nakano, 1996). (Pagliardini et al., 2023) and (Lee et al., 2023) are proposed to explicitly improve the prediction diversity on target data during training. However, target data are typically unavailable during training. To address this limitation, (Rame et al., 2022) and (Wortsman et al., 2022) are proposed to learn a collection of diverse models by varying their learning procedures such as hyperparameters and data augmentations. However, these methods would inevitably produce redundant models, and uniformly ensembling all the models will hurt predictive diversity, leading to compromised performance.

Ensemble pruning. Ensemble pruning, also known as ensemble selection or ensemble thinning, offers a valuable solution to address the limitations of ensemble methods, which often demand extensive memory and processing resources due to the number of individual learners in the ensemble (Liu et al., 2014; Kokkinos & Margaritis, 2015; Zhang et al., 2017). The goal of ensemble pruning is to enhance ensemble generalization performance while reducing its size (Margineantu & Dietterich, 1997). However, selecting the optimal sub-ensemble with superior generalization capabilities is non-trivial, often involving exponential computational complexity (Martinez-Munoz &

Suárez, 2007). Existing ensemble pruning methods can be broadly categorized into three primary families: ranking-based, clustering-based, and optimization-based approaches. Ranking-based pruning methods involve sorting the ensemble learners based on various evaluation criteria and selecting the top performers (Zhang et al., 2019; Ahmed et al., 2017). Clustering-based pruning methods identify groups of learners that make similar predictions and prune these groups individually (Tsoumakas et al., 2004; Cavalcanti et al., 2016). Optimization-based pruning methods leverage various objectives to optimize and identify subsets expected to exhibit satisfactory generalization performance (Zhou & Tang, 2003; Partalas et al., 2012), necessitating the use of optimization algorithms to manage the computational complexity associated with exhaustive searches (Zeng et al., 2014). Despite the effectiveness of these ensemble pruning strategies in in-distribution scenarios, their adaptation to prune models for OoD generalization remains a challenging and open research question.

7. Conclusion

We proposed a novel ensemble pruning framework to improve out-of-distribution generalization. The key insight is to construct an ensemble topology graph that captures predictive relationships between models. This topology is incorporated into a combinatorial optimization problem to jointly optimize for diversity on test data and accuracy on validation data. Through extensive experiments on common benchmarks, the method demonstrates consistent gains over baseline ensembling techniques as well as state-of-the-art domain generalization algorithms. Experimental results showcase significant advancements on challenging multi-source and single-source generalization tasks. A limitation of our approach is the empirical selection of ensemble size rather than a principled theoretical guideline. Future directions include further analysis to automatically determine the optimal number of models for retention.

Acknowledgements

This work is supported by the National Science Foundation through the Faculty Early Career Development (NSF CAREER) Award and the Department of Defense under the Defense Established Program to Stimulate Competitive Research (DoD DEPSCoR) Award.

Impact Statement

This paper presents work whose goal is to advance the field of Machine Learning. There are many potential societal consequences of our work, none which we feel must be specifically highlighted here.

References

Ahmed, M. A. O., Didaci, L., Lavi, B., and Fumera, G. Using diversity for classifier ensemble pruning: an empirical investigation. *Theoretical and Applied Informatics*, 1(29), 2017.

Aksela, M. Comparison of classifier selection methods for improving committee performance. In *International Workshop on Multiple Classifier Systems*, pp. 84–93. Springer, 2003.

Arjovsky, M., Bottou, L., Gulrajani, I., and Lopez-Paz, D. Invariant risk minimization. *arXiv preprint arXiv:1907.02893*, 2019.

Beery, S., Van Horn, G., and Perona, P. Recognition in terra incognita. In *Proceedings of the European conference on computer vision*, pp. 456–473, 2018.

Bian, Y., Wang, Y., Yao, Y., and Chen, H. Ensemble pruning based on objection maximization with a general distributed framework. *IEEE transactions on neural networks and learning systems*, 31(9):3766–3774, 2019.

Breiman, L. Random forests. *Machine learning*, 45:5–32, 2001.

Caruana, R., Niculescu-Mizil, A., Crew, G., and Ksikes, A. Ensemble selection from libraries of models. In *Proceedings of the twenty-first international conference on Machine learning*, pp. 18, 2004.

Cavalcanti, G. D., Oliveira, L. S., Moura, T. J., and Carvalho, G. V. Combining diversity measures for ensemble pruning. *Pattern Recognition Letters*, 74:38–45, 2016.

Cha, J., Chun, S., Lee, K., Cho, H.-C., Park, S., Lee, Y., and Park, S. Swad: Domain generalization by seeking flat minima. *Advances in Neural Information Processing Systems*, 34:22405–22418, 2021.

Dong, X., Yu, Z., Cao, W., Shi, Y., and Ma, Q. A survey on ensemble learning. *Frontiers of Computer Science*, 14: 241–258, 2020.

et al, Z. A spectral clustering based ensemble pruning approach. *Neurocomputing*, 2014.

Fang, C., Xu, Y., and Rockmore, D. N. Unbiased metric learning: On the utilization of multiple datasets and web images for softening bias. In *Proceedings of the IEEE International Conference on Computer Vision*, pp. 1657–1664, 2013.

Fisher, R. A. On the mathematical foundations of theoretical statistics. *Philosophical transactions of the Royal Society of London. Series A, containing papers of a mathematical or physical character*, 222(594-604):309–368, 1922.

Geman, S., Bienenstock, E., and Doursat, R. Neural networks and the bias/variance dilemma. *Neural computation*, 4(1):1–58, 1992.

Gulrajani, I. and Lopez-Paz, D. In search of lost domain generalization. In *International Conference on Learning Representations*, 2020.

Guo, L. and Boukir, S. Margin-based ordered aggregation for ensemble pruning. *Pattern Recognition Letters*, 34(6): 603–609, 2013.

He, K., Zhang, X., Ren, S., and Sun, J. Deep Residual Learning for Image Recognition. In *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 770–778, 2016.

He, Y.-X., Wu, Y.-C., Qian, C., and Zhou, Z.-H. Margin distribution and structural diversity guided ensemble pruning. *Machine Learning*, pp. 1–23, 2024.

Huang, Z., Wang, H., Xing, E. P., and Huang, D. Self-challenging improves cross-domain generalization. In *Proceedings of the European Conference on Computer Vision*, pp. 124–140. Springer, 2020.

Koh, P. W., Sagawa, S., Marklund, H., Xie, S. M., Zhang, M., Balsubramani, A., Hu, W., Yasunaga, M., Phillips, R. L., Gao, I., et al. Wilds: A benchmark of in-the-wild distribution shifts. In *International Conference on Machine Learning*, pp. 5637–5664. PMLR, 2021.

Kokkinos, Y. and Margaritis, K. G. Confidence ratio affinity propagation in ensemble selection of neural network classifiers for distributed privacy-preserving data mining. *Neurocomputing*, 150:513–528, 2015.

Lakshminarayanan, B., Pritzel, A., and Blundell, C. Simple and scalable predictive uncertainty estimation using deep ensembles. *Advances in Neural Information Processing Systems*, 2017.

Lee, Y., Yao, H., and Finn, C. Diversify and disambiguate: Out-of-distribution robustness via disagreement. In *International Conference on Learning Representations*, 2023.

Li, N., Yu, Y., and Zhou, Z.-H. Diversity regularized ensemble pruning. In *European Conference on Machine Learning and Knowledge Discovery in Databases*, pp. 330–345. Springer, 2012.

Li, T., Qiao, F., Ma, M., and Peng, X. Are data-driven explanations robust against out-of-distribution data? In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 3821–3831, 2023.

Li, Z., Ren, K., Jiang, X., Shen, Y., Zhang, H., and Li, D. Simple: Specialized model-sample matching for domain generalization. In *International Conference on Learning Representations*, 2022.

Lin, Y., Tan, L., Hao, Y., Wong, H., Dong, H., Zhang, W., Yang, Y., and Zhang, T. Spurious feature diversification improves out-of-distribution generalization. In *International Conference on Learning Representations*, 2024.

Liu, Z., Dai, Q., and Liu, N. Ensemble selection by grasp. *Applied intelligence*, 41:128–144, 2014.

Ma, M., Li, T., and Peng, X. Beyond the federation: Topology-aware federated learning for generalization to unseen clients. In *International Conference on Machine Learning*, 2024.

Maltese, J., Ombuki-Berman, B. M., and Engelbrecht, A. P. Pareto-based many-objective optimization using knee points. In *2016 IEEE congress on evolutionary computation (CEC)*, pp. 3678–3686. IEEE, 2016.

Margineantu, D. D. and Dietterich, T. G. Pruning adaptive boosting. In *International Conference on Machine Learning*, volume 97, pp. 211–218. Citeseer, 1997.

Martinez-Munoz, G. and Suárez, A. Using boosting to prune bagging ensembles. *Pattern Recognition Letters*, 28(1):156–165, 2007.

Martinez-Munoz, G., Hernández-Lobato, D., and Suárez, A. An analysis of ensemble pruning techniques based on ordered aggregation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 31(2):245–259, 2008.

Matena, M. S. and Raffel, C. A. Merging models with fisher-weighted averaging. *Advances in Neural Information Processing Systems*, 35:17703–17716, 2022.

Miller, J. P., Taori, R., Raghunathan, A., Sagawa, S., Koh, P. W., Shankar, V., Liang, P., Carmon, Y., and Schmidt, L. Accuracy on the line: on the strong correlation between out-of-distribution and in-distribution generalization. In *International Conference on Machine Learning*, 2021.

Nam, H., Lee, H., Park, J., Yoon, W., and Yoo, D. Reducing domain gap by reducing style bias. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 8690–8699, 2021.

Onan, A., Korukoğlu, S., and Bulut, H. A hybrid ensemble pruning approach based on consensus clustering and multi-objective evolutionary algorithm for sentiment classification. *Information Processing & Management*, 53(4): 814–833, 2017.

Pagliardini, M., Jaggi, M., Fleuret, F., and Karimireddy, S. P. Agree to disagree: Diversity through disagreement for better transferability. In *International Conference on Learning Representations*, 2023.

Partalas, I., Tsoumakas, G., and Vlahavas, I. A study on greedy algorithms for ensemble pruning. *Aristotle University of Thessaloniki, Thessaloniki, Greece*, 2012.

Peng, X., Qiao, F., and Zhao, L. Out-of-domain generalization from a single source: An uncertainty quantification approach. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 46(03):1775–1787, 2024.

Qian, C., Yu, Y., and Zhou, Z.-H. Subset selection by pareto optimization. *Advances in Neural Information Processing Systems*, 28, 2015.

Qiao, F. and Peng, X. Uncertainty-guided model generalization to unseen domains. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2021.

Qiao, F. and Peng, X. Topology-aware robust optimization for out-of-distribution generalization. In *International Conference on Learning Representations*, 2023.

Qiao, F., Zhao, L., and Peng, X. Learning to learn single domain generalization. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2020.

Rame, A., Kirchmeyer, M., Rahier, T., Rakotomamonjy, A., Gallinari, P., and Cord, M. Diverse weight averaging for out-of-distribution generalization. *Advances in Neural Information Processing Systems*, 35:10821–10836, 2022.

Ramé, A., Ahuja, K., Zhang, J., Cord, M., Bottou, L., and Lopez-Paz, D. Model ratatouille: Recycling diverse models for out-of-distribution generalization. In *International Conference on Machine Learning*, 2023.

Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., Huang, Z., Karpathy, A., Khosla, A., Bernstein, M., et al. Imagenet large scale visual recognition challenge. *International Journal of Computer Vision*, 115(3): 211–252, 2015.

Sagawa, S., Koh, P. W., Hashimoto, T. B., and Liang, P. Distributionally robust neural networks. In *International Conference on Learning Representations*, 2019.

Sun, Y., Wang, X., Liu, Z., Miller, J., Efros, A., and Hardt, M. Test-time training with self-supervision for generalization under distribution shifts. In *International Conference on Machine Learning*, pp. 9229–9248. PMLR, 2020.

Tsoumakas, G., Katakis, I., and Vlahavas, I. Effective voting of heterogeneous classifiers. In *European Conference on Machine Learning*, pp. 465–476. Springer, 2004.

Tsoumakas, G., Partalas, I., and Vlahavas, I. An ensemble pruning primer. *Applications of supervised and unsupervised ensemble methods*, pp. 1–13, 2009.

Ueda, N. and Nakano, R. Generalization error of ensemble estimators. In *Proceedings of International Conference on Neural Networks*, volume 1, pp. 90–95. IEEE, 1996.

Vapnik, V. *The nature of statistical learning theory*. Springer science & business media, 1999.

Wiles, O., Gowal, S., Stimberg, F., Rebuffi, S.-A., Ktena, I., Dvijotham, K. D., and Cemgil, A. T. A fine-grained analysis on distribution shift. In *International Conference on Learning Representations*, 2021.

Wood, D., Mu, T., Webb, A. M., Reeve, H. W., Lujan, M., and Brown, G. A unified theory of diversity in ensemble learning. *Journal of Machine Learning Research*, 24 (359):1–49, 2023.

Wortsman, M., Ilharco, G., Gadre, S. Y., Roelofs, R., Gontijo-Lopes, R., Morcos, A. S., Namkoong, H., Farhadi, A., Carmon, Y., Kornblith, S., et al. Model soups: averaging weights of multiple fine-tuned models improves accuracy without increasing inference time. In *International Conference on Machine Learning*, pp. 23965–23998. PMLR, 2022.

Yang, M., Zhang, Y., Fang, Z., Du, Y., Liu, F., Ton, J.-F., Wang, J., and Wang, J. Invariant learning via probability of sufficient and necessary causes. *Advances in Neural Information Processing Systems*, 36, 2023.

Zeng, X., Wong, D. F., Chao, L. S., et al. Constructing better classifier ensemble based on weighted accuracy and diversity measure. *The Scientific World Journal*, 2014, 2014.

Zhang, C.-X., Zhang, J.-S., and Yin, Q.-Y. A ranking-based strategy to prune variable selection ensembles. *Knowledge-Based Systems*, 125:13–25, 2017.

Zhang, H., Song, Y., Jiang, B., Chen, B., Shan, G., et al. Two-stage bagging pruning for reducing the ensemble size and improving the classification performance. *Mathematical Problems in Engineering*, 2019, 2019.

Zhang, Y., Burer, S., Nick Street, W., Bennett, K. P., and Parrado-Hernández, E. Ensemble pruning via semi-definite programming. *Journal of Machine Learning Research*, 7(7), 2006.

Zhou, Z.-H. and Tang, W. Selective ensemble of decision trees. In *Rough Sets, Fuzzy Sets, Data Mining, and Granular Computing: 9th International Conference, RSFDGrC 2003, Chongqing, China, May 26–29, 2003 Proceedings 9*, pp. 476–483. Springer, 2003.

Zhou, Z.-H., Wu, J., and Tang, W. Ensembling neural networks: many could be better than all. *Artificial intelligence*, 137(1-2):239–263, 2002.

A. Proofs

We prove the step-by-step proof for Lem. 4.1 (Diversity Promotion) and Thm. 4.4 (Generalization Error of the Pruned Ensemble) in the main paper.

Lemma 4.1 (Diversity Promotion). Let $S \subseteq V$ be the pruned ensemble with the size of K, obtained by solving the optimization problem in Eq. 6. Define the average pairwise similarity among all models in V as:

$$W_V = \frac{2}{N(N-1)} \sum_{i<j} W_{ij}, \tag{14}$$

and the average pairwise similarity among models in the pruned ensemble S as:

$$W_S = \frac{2}{K(K-1)} \sum_{i<j} W_{ij} z_i z_j. \tag{15}$$

Then, the following statement holds:

$$1 - W_S \geq 1 - W_V. \tag{16}$$

Proof. Let $z^* = [z_1^*, z_2^*, \ldots, z_N^*]$ be the optimal solution obtained by solving the optimization problem in Eq. 7. The objective function value of the optimal solution is:

$$\begin{aligned} J(S^*) &= \sum_{i<j} (1 - W_{ij}) z_i^* z_j^* \\ &= \sum_{i=1}^{N} \sum_{j=i+1}^{N} (1 - W_{ij}) z_i^* z_j^* \\ &= \frac{K(K-1)}{2} (1 - W_S). \quad (\text{using } Eq.\ 15 \text{ and the constraint } \sum_{i=1}^{N} z_i^* = K) \end{aligned}$$

Similarly, the objective function value of selecting all models in V is:

$$\begin{aligned} J(V) &= \sum_{i<j} (1 - W_{ij}) \\ &= \sum_{i=1}^{N} \sum_{j=i+1}^{N} (1 - W_{ij}) \\ &= \frac{N(N-1)}{2} (1 - W_V). \quad (\text{using } Eq.\ 14) \end{aligned}$$

Since S^* is the optimal solution, $J(S^*) \geq J(S)$ for any $S \subseteq V$ with $|S| = K$. In particular, let S' be a random subset of V with size K. Then, we have:

$$\begin{aligned} E[J(S')] &\leq J(S^*) \\ \Rightarrow \frac{K(K-1)}{2} E[1 - W_{S'}] &\leq \frac{K(K-1)}{2} (1 - W_S) \\ \Rightarrow 1 - W_S &\geq E[1 - W_{S'}]. \end{aligned}$$

Note that $E[W_{S'}] = W_V$ since S' is a random subset of V. Therefore, we have:

$$1 - W_S \geq 1 - W_V.$$

This completes the proof. □

Proposition A.1 (Generalised Ambiguity Decomposition, Adapted from (Wood et al., 2023)). *Assuming loss ℓ admits a bias-variance decomposition then, for an ensemble $\left\{\hat{Y}_i\right\}_{i=1}^{K}$, the ambiguity decomposition is formulated as:*

$$\underbrace{\ell(\overline{Y}, Y)}_{\text{Ensemble loss}} = \underbrace{\frac{1}{K}\sum_{i=1}^{K}\ell\left(\hat{Y}_i, Y\right)}_{\text{Average loss}} - \underbrace{\frac{1}{K}\sum_{i=1}^{K}\ell\left(\hat{Y}_i, \overline{Y}\right)}_{\text{Ambiguity}}. \tag{17}$$

Theorem 4.4 (Generalization Error of the Pruned Ensemble. Adapted from (Wood et al., 2023)). Given a set of model predictions $\{\hat{Y}_i\}_{i=1}^{K}$ and a loss function ℓ, assuming a bias-variance decomposition holds in Def. 4.2, the following decomposition also holds:

$$\underbrace{\mathbb{E}_D[\mathbb{E}_{X,Y}[\ell(\overline{Y}, Y)]]}_{\text{Expected ensemble risk}} = \mathbb{E}_X[\underbrace{\mathbb{E}_{Y|X}[\ell(Y^*, Y)]}_{\text{Noise}} + \underbrace{\frac{1}{K}\sum_{i=1}^{K}\ell(\tilde{Y}_i, Y^*)}_{\text{Average bias}} + \underbrace{\frac{1}{K}\sum_{i=1}^{K}\mathbb{E}_D[\ell(\hat{Y}_i, \tilde{Y}_i)]}_{\text{Average variance}} - \underbrace{\mathbb{E}_D[\frac{1}{K}\sum_{i=1}^{K}\ell(\hat{Y}_i, \overline{Y})]}_{\text{Diversity}}], \tag{18}$$

where $\tilde{Y}_i \overset{\text{def}}{=} \arg\min_{Y\in\mathcal{Y}}\mathbb{E}_D[\ell(\hat{Y}_i, Y)]$. Eq. 18 decomposes the expected ensemble risk into the noise, the average bias, the average variance, and the diversity.

Proof. After taking the expected risk of $\overline{Y}$ and applying Prop. A.1, we can obtain:

$$\mathbb{E}_D\left[\mathbb{E}_{X,Y}[\ell(\overline{Y}, Y)]\right] = \mathbb{E}_D\left[\mathbb{E}_{X,Y}\left[\frac{1}{K}\sum_{i=1}^{K}\ell\left(\hat{Y}_i, Y\right)\right]\right] - \mathbb{E}_D\left[\mathbb{E}_{X,Y}\left[\frac{1}{K}\sum_{i=1}^{K}\ell\left(\hat{Y}_i, \overline{Y}\right)\right]\right]. \tag{19}$$

Then we apply Def. 4.2 to the first term on the right hand side of Eq. 19:

$$\begin{aligned}&\mathbb{E}_D\left[\mathbb{E}_{X,Y}\left[\frac{1}{K}\sum_{i=1}^{K}\ell\left(\hat{Y}_i, Y\right)\right]\right] = \\ &\mathbb{E}_X\left[\mathbb{E}_{Y|X}\left[\ell\left(Y^*, Y\right)\right] + \frac{1}{K}\sum_{i=1}^{K}\ell\left(\tilde{Y}_i, Y^*\right) + \frac{1}{K}\sum_{i=1}^{K}\mathbb{E}_D\left[\ell\left(\hat{Y}_i, \tilde{Y}_i\right)\right]\right].\end{aligned} \tag{20}$$

We can complete the proof by plugging Eq. 20 into Eq. 19. □

B. Additional Results

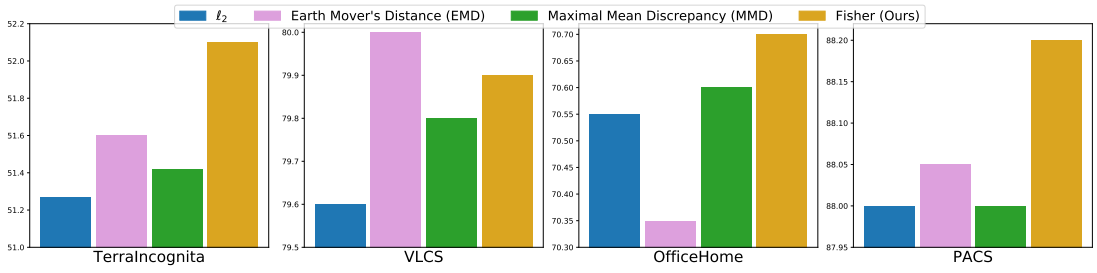

Figure 6. Accuracy (%) of alternative distance metrics for model topology construction. We compare with several common distance metrics including ℓ_2, Earth Mover's Distance (EMD), and Maximum Mean Discrepancy (MMD). Results show Fisher is only slightly worse than EMD on *VLCS* and outperforms all the other metrics on other datasets. We attribute the effectiveness to Fisher's sensitivity in capturing the parameter-output relationships (Matena & Raffel, 2022).

Alternative distance metrics for ensemble topology construction. We compare with several common distance metrics including ℓ_2, Earth Mover's Distance (EMD), and Maximum Mean Discrepancy (MMD). The results are shown in Fig. 6.

Models	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
z	0.37	1.27e-21	0.50	1.48e-21	1.00	1.00	1.00	1.00	6.15e-22	0.68	1.00	0.89e-3	1.00	1.00	1.44e-21	0.44	1.26e-21	1.00	2.45e-22	1.39e-21

Table 6. Values of z obtained by solving the problem in Eq. 7. The values of almost half of the models are smaller than 1e-3. Consequently, utilizing these values for a weighted ensemble may not be informative, as the contributions of these models would be negligible.

Results show Fisher is only slightly worse than EMD on *VLCS* and outperforms all the other metrics on other datasets. We attribute the effectiveness to Fisher's sensitivity in capturing the parameter-output relationships (Matena & Raffel, 2022).

Weighted ensemble (soft version) vs. Pruning. We show the values of z obtained by solving the problem in Eq. 7. The results on the *Terra Incognita* dataset are shown in Tab. 6. The table reveals that the values of almost half of the models are smaller than 1e-3. Consequently, utilizing these values for a weighted ensemble may not be informative, as the contributions of these models would be negligible. In this case, removing these models can significantly improve inference efficiency while maintaining or even improving OoD generalization performance. Binary selection allows us to directly control the size of the pruned ensemble, which is crucial for computational efficiency.