Feed Two Birds with One Scone: Exploiting Wild Data for Both Out-of-Distribution Generalization and Detection

Haoyue Bai [1] **Gregory Canal** [2] **Xuefeng Du** [1] **Jeongyeol Kwon** [3] **Robert Nowak** [3] **Yixuan Li** [1]

Abstract

Modern machine learning models deployed in the wild can encounter both covariate and semantic shifts, giving rise to the problems of out-of-distribution (OOD) generalization and OOD detection respectively. While both problems have received significant research attention lately, they have been pursued independently. This may not be surprising, since the two tasks have seemingly conflicting goals. This paper provides a new unified approach that is capable of simultaneously generalizing to covariate shifts while robustly detecting semantic shifts. We propose a margin-based learning framework that exploits freely available unlabeled data in the wild that captures the environmental test-time OOD distributions under both covariate and semantic shifts. We show both empirically and theoretically that the proposed margin constraint is the key to achieving both OOD generalization and detection. Extensive experiments show the superiority of our framework, outperforming competitive baselines that specialize in either OOD generalization or OOD detection. Code is publicly available at https://github.com/deeplearning-wisc/scone.

1. Introduction

Modern machine learning models deployed in the wild can encounter different types of distributional shifts. Taking autonomous driving as an example, a model trained on in-distribution (ID) data with sunny weather (Figure 1, left) may experience a *covariate shift* due to snowy weather (Figure 1, middle). Under such a covariate shift, a model is expected to generalize to the out-of-distribution (OOD) data—correctly predicting the sample into one of the known classes (e.g., car), despite the shift. Additionally, the model may encounter a *semantic shift*, where samples are from unknown classes (e.g., deer) that the model has not been exposed to during training (Figure 1, right). Such semantic OOD data should be rejected instead of being blindly predicted as a known class.

These distributional shift scenarios give rise to the importance of two problems: OOD generalization, which focuses on the covariate shift problem (Gulrajani & Lopez-Paz, 2020; Koh et al., 2021; Ye et al., 2022), and OOD detection, which targets semantic shift (Hendrycks & Gimpel, 2017; Liu et al., 2020; Yang et al., 2021). Both problems have received increasing research attention lately, albeit have been pursued independently; as a result, existing methods are highly specialized in one task, but not capable of handling both simultaneously. This has largely impeded the wider adoption of OOD algorithms in real-world environments, which often present heterogeneous data shifts. A critical yet underexplored question thus arises:

Can we devise a unified learning framework for both OOD generalization and OOD detection?

In this paper, we bridge the gap between OOD generalization and OOD detection, in one coherent framework. Our driving idea is to exploit unlabeled wild data naturally arising in the model's operating environment, turning the OOD threat into valuable learning resources instead. Wild data arises naturally *for free* upon deploying any machine learning classifier in its respective environment, and has been largely overlooked for OOD learning purposes. Specifically, we consider a generalized characterization of the wild data, which can be modeled as a mixed composition of three data distributions:

$$\mathbb{P}_{\text{wild}} := (1-\pi_s-\pi_c)\mathbb{P}_{\text{in}} + \pi_c\mathbb{P}_{\text{out}}^{\text{covariate}} + \pi_s\mathbb{P}_{\text{out}}^{\text{semantic}},$$

where $\mathbb{P}_{\text{in}}$, $\mathbb{P}_{\text{out}}^{\text{covariate}}$ and $\mathbb{P}_{\text{out}}^{\text{semantic}}$ denote the marginal distributions of ID, covariate-shifted OOD and semantic-shifted OOD data respectively. Such wild data is available in abundance, does not require any human annotation, and importantly, contains the *true* test time OOD distributions under

[1]Department of Computer Sciences, University of Wisconsin, Madison [2]Institute for Foundations of Data Science, University of Wisconsin, Madison [3]Department of Electrical and Computer Engineering, University of Wisconsin, Madison. Correspondence to: Yixuan Li <sharonli@cs.wisc.edu>.

Proceedings of the 40^{th} International Conference on Machine Learning, Honolulu, Hawaii, USA. PMLR 202, 2023.

Figure 1. Illustration of three types of data that can organically arise when deploying models in the open world: (1) in-distribution (ID) data (e.g., car on a sunny day), (2) covariate-shifted OOD data (e.g., car in the snow), and (3) semantic-shifted OOD data (e.g., a deer). Our framework enables leveraging the wild mixture (containing all three types of data) for OOD generalization and OOD detection.

both covariate and semantic shifts. Thus, our distributional model above offers strong generality and practicality, compared to previous works that primarily consider the semantic shift in the wild data (Katz-Samuels et al., 2022). Despite the promise, learning from such heterogeneous data is technically challenging due to the lack of clear membership (ID, Covariate-OOD, Semantic-OOD) for samples drawn from the wild data distribution $\mathbb{P}_{\text{wild}}$.

To tackle this challenge, we formulate a new learning framework SCONE—**S**emantic and **C**ovariate **O**ut-of-distribution Lear**N**ing via **E**nergy Margins. SCONE jointly learns a robust multi-class classifier that generalizes to covariate-OOD data, and a reliable OOD detector that detects semantic-OOD data. Our key idea is to explicitly optimize for a binary classifier based on the energy function, classifying as many samples as possible from $\mathbb{P}_{\text{wild}}$ as OUT (with positive energy), subject to two constraints (i) the ID data has energy smaller than a negative margin value, and (ii) the multi-class classification model must maintain high accuracy. We show both theoretically (Section 3.3) and empirically (Section 4) that the margin constraint is the key to the success of our algorithm. Intuitively, enforcing an energy margin on ID data has the effect of also lowering the energy of nearby covariate-OOD points, which are semantically related to ID points. Since lower energy increases the value of the classifier logits, the covariate-OOD points then enjoy an increased logit in their correct classes, leading to stronger OOD generalization.

Extensive experiments confirm that SCONE can effectively improve both OOD generalization and detection performance. Compared to the most related baseline WOODS (Katz-Samuels et al., 2022), our method can substantially improve the OOD classification accuracy from 52.76% to 84.69% on covariate shifted CIFAR-10 data—a direct **31.93**$\%$ improvement. Our key contributions are:

- To the best of our knowledge, we are among the first works that utilize wild data to jointly tackle two tasks of OOD generalization and OOD detection in one framework. Our problem formulation offers strong generality and practicality for real-world applications.
- We propose a margin-based learning framework that exploits freely available, unlabeled data in the wild to solve our problem. We model wild data as a comprehensive mixture of ID samples, covariate OOD, and semantic OOD data.
- We perform extensive experiments and ablations, which demonstrate the efficacy of our method. We show that SCONE demonstrates overall strong performance in both OOD generalization and detection, outperforming baselines that specialize in one or the other.

2. Problem Setup

Labeled in-distribution data. Let $\mathcal{X} = \mathbb{R}^d$ denote the input space and $\mathcal{Y} = \{1, \dots, K\}$ denote the label space. We assume access to a labeled training set $\mathcal{D}_{\text{in}}^{\text{train}} = \{(\mathbf{x}_i, y_i)\}_{i=1}^n$, drawn *i.i.d.* from the joint data distribution $\mathbb{P}_{\mathcal{X}\mathcal{Y}}$. Let $\mathbb{P}_{\text{in}}$ denote the marginal distribution on $\mathcal{X}$, which is also referred to as the *in-distribution*. Let $f_\theta : \mathcal{X} \mapsto \mathbb{R}^K$ denote a function for the classification task, which predicts the label of an input sample $\mathbf{x}$ as $\widehat{y}(f_\theta(\mathbf{x})) \coloneqq \arg\max_y f_\theta^{(y)}(\mathbf{x})$, where $f_\theta^{(y)}(\mathbf{x})$ denotes the y-th element of $f_\theta(\mathbf{x})$, corresponding to label y.

Unlabeled wild data. Trained on the ID data, the classifier f_θ deployed into the wild can encounter various distributional shifts (see Figure 1). To model the realistic environment, we consider the following generalized characterization of the wild data:

$$\mathbb{P}_{\text{wild}} \coloneqq (1 - \pi_c - \pi_s)\mathbb{P}_{\text{in}} + \pi_c \mathbb{P}_{\text{out}}^{\text{covariate}} + \pi_s \mathbb{P}_{\text{out}}^{\text{semantic}}, \quad (1)$$

where $\pi_c, \pi_s, \pi_c + \pi_s \in [0, 1]$. Our mathematical formulation thus fully encapsulates all three possible distributions that the deployed model may encounter in practice:

- **In-distribution** $\mathbb{P}_{\text{in}}$ is the marginal distribution of the labeled data.
- **Covariate OOD** $\mathbb{P}_{\text{out}}^{\text{covariate}}$ is the marginal distribution of $\mathbb{P}_{\mathcal{X}'\mathcal{Y}}$ on $\mathcal{X}'$, where the joint distribution has the same label space as the training data, yet the input space undergoes shifting in style and domain. This is relevant for *OOD generalization*.

Method	OOD Accuracy (MNIST-C) (OOD generalization) ↑	ID Accuracy (MNIST) (ID generalization) ↑	FPR95 (OOD detection) ↓	AUROC (OOD detection) ↑
WOODS (Katz-Samuels et al., 2022)	88.10%	97.88%	0.194%	99.88%
Ours	**96.51**%	97.79%	0.017%	99.99%

Table 1. WOODS (Katz-Samuels et al., 2022) displays limiting OOD generalization performance. For experiments, we use 25,000 samples from the MNIST dataset as ID, and a wild mixture dataset consisting of FashionMNIST as semantic-OOD data (π_s = 0.4) and MNIST-C (Mu & Gilmer, 2019) — a covariate shifted version of MNIST — as covariate-OOD data ($\pi_c = 0.3$).

- **Semantic OOD** $\mathbb{P}_{\text{out}}^{\text{semantic}}$: wild data that does not belong to any known categories $\mathcal{Y} = \{1, 2, ..., K\}$, and therefore should not be predicted by the model. This is relevant for *OOD detection*.

Learning goal. Our learning framework revolves around building an OOD detector $g_\theta \colon \mathcal{X} \to \{\text{IN}, \text{OUT}\}$ and multi-class classifier f_θ by leveraging data from both $\mathbb{P}_{\text{in}}$ and $\mathbb{P}_{\text{wild}}$. The OOD detector g_θ should predict semantic OOD data as OUT and otherwise predict as IN[1]. We notate g_θ and f_θ as sharing parameters θ to indicate the fact that these functions may share neural network parameters. In evaluating our model, we are interested in the following measurements:

$$(1)\ \uparrow \text{ID-Acc}(f_\theta) := \mathbb{E}_{(\mathbf{x},y)\sim\mathbb{P}_{\mathcal{X}\mathcal{Y}}}(\mathbb{1}\{\widehat{y}(f_\theta(\mathbf{x})) = y\}),$$
$$(2)\ \uparrow \text{OOD-Acc}(f_\theta) := \mathbb{E}_{(\mathbf{x},y)\sim\mathbb{P}_{\text{out}}^{\text{covariate}}}(\mathbb{1}\{\widehat{y}(f_\theta(\mathbf{x})) = y\}),$$
$$(3)\ \downarrow \text{FPR}(g_\theta) := \mathbb{E}_{\mathbf{x}\sim\mathbb{P}_{\text{out}}^{\text{semantic}}}(\mathbb{1}\{g_\theta(\mathbf{x}) = \text{IN}\}),$$

where $\mathbb{1}\{\cdot\}$ is the indicator function and the arrows indicate higher/lower is better. ID-Acc, OOD-Acc, and FPR jointly capture the **(1)** ID generalization, **(2)** OOD generalization, and **(3)** OOD detection performance, respectively. In the context of OOD detection, ID samples are considered positive, and FPR means false positive rate.

3. Methodology

In this section, we present a unified learning framework that enables performing both OOD generalization and OOD detection, by way of exploiting unlabeled data in the wild. Our framework offers substantial advantages over the counterpart approaches that rely only on the ID data, and naturally suits many applications where machine learning models are deployed in the open world. We start with preliminaries to lay the necessary context (Section 3.1), followed by our proposed method (Section 3.2) and theory (Section 3.3).

3.1. Preliminaries

Katz-Samuels et al. (2022) proposed WOODS to tackle the OOD detection problem via unlabeled wild data, which consists of the ID and semantically shifted OOD data $\mathbb{P}_{\text{wild}} := (1-\pi)\mathbb{P}_{\text{in}} + \pi\mathbb{P}_{\text{out}}^{\text{semantic}}$. The crucial difference between our work and WOODS is whether the wild mixture data contains covariate-shifted data, which introduces new challenges not considered in prior work. As we show in this work, our formulation uniquely enables both OOD generalization and OOD detection, in one coherent framework.

As preliminaries, WOODS minimizes the error of declaring data from $\mathbb{P}_{\text{wild}}$ as ID, subject to *(i)* the error of declaring an ID point as OOD is at most a fixed threshold α, and *(ii)* the multi-class classification model meets some error threshold τ. Mathematically, this can be formalized as a constrained optimization problem:

$$\begin{aligned} &\text{argmin}_\theta\ \mathbb{E}_{\mathbf{x}\sim\mathbb{P}_{\text{wild}}}(\mathbb{1}\{g_\theta(\mathbf{x}) = \text{IN}\}) \\ &\text{s.t. } \mathbb{E}_{\mathbf{x}\sim\mathbb{P}_{\text{in}}}(\mathbb{1}\{g_\theta(\mathbf{x}) = \text{OUT}\}) \le \alpha \\ &\quad \mathbb{E}_{(\mathbf{x},y)\sim\mathbb{P}_{\mathcal{X}\mathcal{Y}}}(\mathbb{1}\{\widehat{y}(f_\theta(\mathbf{x})) \neq y\}) \le \tau. \end{aligned} \tag{2}$$

In particular, the OOD detector g_θ is defined based on the level set: $g_\theta(\mathbf{x}) = \text{OUT}$ if $E_\theta(\mathbf{x}) > 0$, where the free energy $E_\theta(\mathbf{x}) := -\log\sum_{j=1}^{K} e^{f_\theta^{(j)}(\mathbf{x})}$ was shown to be an effective OOD score (Liu et al., 2020). ID data tends to have negative energy and vice versa.

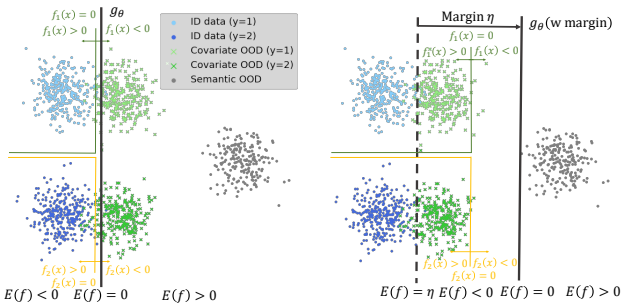
In practice, the objective in (2) can be empirically optimized over *i.i.d.* samples $\widetilde{\mathbf{x}}_1 \dots \widetilde{\mathbf{x}}_m \sim \mathbb{P}_{\text{wild}}$ and $\mathbf{x}_1 \dots \mathbf{x}_n \sim \mathbb{P}_{\text{in}}$ via a tractable relaxation by replacing the $0/1$ loss with a surrogate loss as follows:

$$\begin{aligned} &\text{argmin}_\theta\ \frac{1}{m}\sum_{i=1}^{m} \mathcal{L}_{\text{ood}}(g_\theta(\tilde{\mathbf{x}}_i), \text{IN}) \\ &\text{s.t. } \frac{1}{n}\sum_{j=1}^{n} \mathcal{L}_{\text{ood}}(g_\theta(\mathbf{x}_j), \text{OUT}) \le \alpha \\ &\quad \frac{1}{n}\sum_{j=1}^{n} \mathcal{L}_{\text{cls}}(f_\theta(\mathbf{x}_j), y_j) \le \tau, \end{aligned} \tag{3}$$

where $\mathcal{L}_{\text{ood}}(g_\theta(\mathbf{x}_i), \text{OUT}) = \frac{1}{1+\exp(-w\cdot E_\theta(\mathbf{x}_i))}$ denotes the loss of the binary OOD classifier (where $w \in \mathbb{R}$ is a learnable parameter) and $\mathcal{L}_{\text{cls}}(f_\theta(\mathbf{x}), y)$ is the per-sample cross-entropy (CE) loss for the classification task.

Limitation in OOD generalization performance of WOODS: a case study. Although WOODS can simultaneously learn an OOD detector and an ID classifier, it can perform poorly on the task of OOD generalization. To see

[1]We use OUT to avoid abusing the term OOD. In the context of OOD detection, OUT refers particularly to "outside the semantic space $\mathcal{Y}$". Hence covariate-OOD falls into the IN category, in the semantic sense.



(a) No margin ($\eta = 0$) (b) Energy margin ($\eta < 0$)

Figure 2. Illustration of the impact of energy margin η on the placement of the OOD detection boundary.

this, we investigate the efficacy of using the constrained optimization in Equation (3) directly for our new problem setting described in Section 2. To simulate the wild data $\mathbb{P}_{\text{wild}}$, we mix a subset of MNIST training data (as $\mathbb{P}_{\text{in}}$) with MNIST-C (Mu & Gilmer, 2019) data — a covariate-shifted version of MNIST — as $\mathbb{P}_{\text{out}}^{\text{covariate}}$, and the FashionMNIST dataset (Xiao et al., 2017) as $\mathbb{P}_{\text{out}}^{\text{semantic}}$. We train a version of WOODS on this data and summarize our findings in Table 1: interestingly, we observe a significant generalization gap between classifying ID data (97.88% in accuracy) and covariate-shifted OOD data (88.10% in accuracy). This suggests that the training objective in (Katz-Samuels et al., 2022) is indeed insufficient for the purpose of OOD generalization, despite strong OOD detection performance. This motivates our method which accounts for the task of OOD generalization on covariate-shifted data.

3.2. Proposed Method

Motivation. Before we present our method, an important step is to identify the key reason for the limited OOD generalization performance observed when using WOODS. To better understand the behavior of WOODS, we use a simple two-class example ($K = 2$) to illustrate the decision boundary of its energy-based OOD detector $g_\theta(\mathbf{x}) = \text{OUT}$ if $E_\theta(\mathbf{x}) > 0$ (black vertical line in Figure 2a). Geometrically, since $E_\theta(\mathbf{x})$ is high when $f_\theta^{(1)}$ and $f_\theta^{(2)}$ are both low, the objective in (2) attempts to minimize $f_\theta(\mathbf{x})$ elementwise on points in $\mathbb{P}_{\text{wild}}$. Conversely, since $f_\theta(\mathbf{x})$ is simultaneously being trained to classify ID points (in blue) correctly, it will be optimized towards being "one-hot" on each point in $\mathbb{P}_{\text{in}}$.

With this perspective, our critical insight is that the objective in Equation (2) incentivizes the OOD detector to be as close as possible to the ID data (colored in blue), since it aims to classify as many samples as possible from $\mathbb{P}_{\text{wild}}$ — which includes covariate-shifted data in this setting — as semantic OUT. Therefore, the OOD detection boundary undesirably places the covariate-shifted data (colored in green) on the wrong side — labeling it as semantic OUT — and drives its label distribution towards uniform (since neither $f_\theta^{(1)}$ or $f_\theta^{(2)}$ are "one-hot"), resulting in classification errors. Instead, the OOD detector should ideally place only the semantically shifted OOD data (colored in gray) on the OUT side of the detection boundary.

Proposed margin-based learning objective. Leveraging these insights, we now propose a new learning objective to mitigate this issue. SCONE is motivated by the observation in Figure 2a, where the OOD decision boundary lacks the sufficient margin *w.r.t.* the ID data. Therefore, our key idea is to enforce a sufficient margin between the OOD detector and the ID data.

Our key idea is to enforce the *ID data to have energy smaller than the margin* η (a negative value), while optimizing for the level-set estimation based on the energy function. Given access to samples $\tilde{\mathbf{x}}_1, \ldots, \tilde{\mathbf{x}}_m$ from $\mathbb{P}_{\text{wild}}$, along with labeled ID samples $(\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_n, y_n)$, our proposed optimization is:

$$\begin{aligned} \text{argmin}_\theta \; & \frac{1}{m}\sum_{i=1}^{m} \mathbb{1}\{E_\theta(\tilde{\mathbf{x}}_i) \leq 0\} \\ \text{s.t.} \; & \frac{1}{n}\sum_{j=1}^{n} \mathbb{1}\{E_\theta(\mathbf{x}_j) \geq \eta\} \leq \alpha \\ & \frac{1}{n}\sum_{j=1}^{n} \mathbb{1}\{\widehat{y}(f_\theta(\mathbf{x}_j)) \neq y_j\} \leq \tau, \end{aligned} \tag{4}$$

where η controls the margin of the OOD decision boundary *w.r.t.* the ID data. Note that our learning objective generalizes WOODS (Katz-Samuels et al., 2022), which corresponds to the special case of $\eta = 0$ (i.e., no margin at all).

We illustrate the intuitive effect of this ID energy margin constraint in Figure 2b: by requiring lower energy on ID points with a hard constraint in (4), the OOD detector g_θ is forced to move its zero level-set to the right in order to decrease ID energy. Since decreasing $E_\theta(\mathbf{x})$ directly corresponds to increasing $f_\theta^{(1)}(\mathbf{x})$ and/or $f_\theta^{(2)}(\mathbf{x})$, this is achieved by moving the zero level-sets of $f_\theta^{(1)}(\mathbf{x})$ and $f_\theta^{(2)}(\mathbf{x})$ to the right. Since $f_\theta^{(1)}(\mathbf{x})$ and $f_\theta^{(2)}(\mathbf{x})$ are also constrained to classify ID data correctly, we expect that this level-set shift will still preserve the classifier boundary between ID points, resulting in the geometry visualized in Figure 2b. Crucially, due to the assumed nearness of covariate-shifted points to ID points, these shifts can have the effect of then generalizing the ID classification boundary to $\mathbb{P}_{\text{out}}^{\text{covariate}}$. Although (4) is also attempting to maximize the number of points from $\mathbb{P}_{\text{out}}^{\text{covariate}}$ detected as OUT (i.e., on the right side of the $E_\theta(\mathbf{x}) = 0$ boundary), this is a weaker force than the ID energy margin, since the latter is a hard constraint. To provide some theoretical insights, we next formalize these intuitions for a restricted model architecture.

3.3. Theoretical Insights: Relationship Between OOD Generalization and Energy Margin

We now discuss how SCONE can *improve the classification* accuracy on the covariate-shifted data under some assumptions on the data distribution and model class. Suppose we are given a fixed feature map $\phi\colon \mathbb{R}^d \to \mathbb{R}^p$ to some feature space in $\mathbb{R}^p$, where there exists a constant $\delta > 0$ such that for each covariate-shifted point $\mathbf{x}_c$ with ground-truth label y, there exists a corresponding ID point $\mathbf{x}$, also with label y, satisfying $\|\phi(\mathbf{x}_c) - \phi(\mathbf{x})\|_2 < \delta$. That is, we assume that the covariate-shifted data is close to in-distribution data in the feature space, which is supported empirically in our experiments (Section 4.3). Suppose a classifier $f_\theta\colon \mathbb{R}^p \to \mathbb{R}^K$ is learned on top of $\phi(\mathbf{x})$, using the "idealized" version of our method in (4) (replacing each data point $\mathbf{x}_i, \widetilde{\mathbf{x}}_i$ in (4) with its corresponding feature vector $\phi(\mathbf{x}_i), \phi(\widetilde{\mathbf{x}}_i)$), and consider the two-class case ($K = 2$). Since each classification decision only depends on the difference $f_\theta^{(1)}(\cdot) - f_\theta^{(2)}(\cdot)$, for analytical convenience suppose that we learn this difference directly as $\overline{f}_\theta(\cdot) = f_\theta^{(1)}(\cdot) - f_\theta^{(2)}(\cdot)$ with $\overline{f}_\theta(\cdot) > 0$ corresponding to $y = 1$ and $\overline{f}_\theta(\cdot) < 0$ indicating $y = 2$, which we can accomplish in our framework by fixing $f_\theta^{(2)} = -\frac{1}{2}\overline{f}_\theta$ and $f_\theta^{(1)} = \frac{1}{2}\overline{f}_\theta$.

Suppose further that we set $\alpha = 0$ and $\tau = 0$, such that every ID point is classified correctly and has energy satisfying $E_\theta(\phi(\mathbf{x})) < \eta$. Finally, suppose that $\overline{f}_\theta$ is L-Lipschitz: this is the case for many classifier functions, such as two-layer ReLU networks with bounded variation (Parhi & Nowak, 2022). We then have the following result on the covariate-shifted points (proved in Appendix A):

Proposition 3.1. *Under the above assumptions, if* $\eta < -\log 2 - \frac{1}{2}L\delta$ *then each covariate-shifted point is classified correctly and is detected as semantic* IN.

Implications. This result illustrates that even though our method does not have access to the ground-truth labels of covariate-shifted data during training, we expect that as long as each covariate-shifted data point is "close" to a corresponding ID point, then setting the ID energy threshold η appropriately will result in the covariate-shifted data being classified *correctly*. Intuitively, requiring the ID data points to have lower energy while simultaneously being classified correctly encourages their logit values $f_\theta(\cdot)$ to move further from the classification decision boundary, on the correct side. If f_θ belongs to a regularized function class (e.g., Lipschitz functions), then the logits of the covariate-shifted data will not deviate wildly from their ID counterparts. Combining these two insights, the logits of the covariate-shifted data should also be bounded away from the decision boundary, on the correct side. Importantly, this distance to the boundary is explicitly increased by decreasing η, whereas WOODS does not necessarily ensure this property.

3.4. Enforcing Margin in Practice

Since the $0/1$ loss in (4) is intractable, in a similar manner to WOODS, we replace it with a smooth approximation given by the binary sigmoid loss, yielding the following optimization problem:

$$\begin{aligned} \operatorname{argmin}_{\theta, w\in\mathbb{R}} \ & \frac{1}{m}\sum_{i=1}^{m} \frac{1}{1+\exp(w\cdot E_\theta(\tilde{\mathbf{x}}_i))} \\ \text{s.t.} \ & \frac{1}{n}\sum_{j=1}^{n} \frac{1}{1+\exp(-w\cdot(E_\theta(\mathbf{x}_j)-\eta))} \le \alpha \\ & \frac{1}{n}\sum_{j=1}^{n} \mathcal{L}_{\text{cls}}(f_\theta(\mathbf{x}_j), y_j) \le \tau, \end{aligned} \tag{5}$$

Solving the constrained optimization. We adopt the Augmented Lagrangian method (Hestenes, 1969) to solve our constrained optimization problem with modern neural networks. In short, the constrained optimization problem above is converted into a sequence of unconstrained optimization problems. We refer interested readers to Section 3.2 in Katz-Samuels et al. (2022) for details. We showcase the efficacy of our algorithm on the simple MNIST example in Table 1, where the proposed method improves the OOD generalization accuracy from 88.10% (WOODS) to **96.51%**. Building on this encouraging result, we proceed to comprehensively evaluate our algorithm in the next section.

4. Experiments

In this section, we comprehensively verify the empirical efficacy of SCONE. We first describe the experimental setup (Section 4.1). In Section 4.2, we present results for both OOD generalization and OOD detection, followed by extensive ablations. We provide qualitative analysis that improves the understanding of SCONE in Section 4.3.

4.1. Experimental Setup

Datasets and evaluation metrics. Following the common benchmarks in literature, we use CIFAR-10 (Krizhevsky et al., 2009) as the in-distribution data ($\mathbb{P}_{\text{in}}$). For the covariate-shifted data ($\mathbb{P}_{\text{out}}^{\text{covariate}}$), we use CIFAR-10-C (Hendrycks & Dietterich, 2018) with Gaussian additive noise for our main experiments, and provide ablations in Appendix F on other types of covariate shifts. For semantic-shifted OOD data ($\mathbb{P}_{\text{out}}^{\text{semantic}}$), we use natural image datasets: SVHN (Netzer et al., 2011), Textures (Cimpoi et al., 2014), Places365 (Zhou et al., 2017), LSUN-Crop (Yu et al., 2015), and LSUN-Resize (Yu et al., 2015). Large-scale results on the ImageNet dataset can be found in Section 4.4. Additional results on the PACS dataset (Li et al., 2017) from DomainBed is presented in Appendix E. We provide a detailed description of the datasets in Appendix C.

To simulate the wild data $\mathbb{P}_{\text{wild}}$, we mix a subset of ID

data ($\mathbb{P}_{\text{in}}$) with the covariate shifted dataset ($\mathbb{P}_{\text{out}}^{\text{covariate}}$) under various $\pi_c \in \{0.0, 0.1, 0.2, 0.5, 0.9\}$; we default to $\pi_c = 0.5$ for the main experiment. We keep $\pi_s = 0.1$ to reflect the fact that we expect semantic OOD shifts ($\mathbb{P}_{\text{out}}^{\text{semantic}}$) to be encountered less frequently, which is the same value as used in (Katz-Samuels et al., 2022). We split ID datasets into two halves: we use 50% as ID training data and 50% for creating the mixture data.

In each training iteration, we simulate the mixture data as follows. For the ID dataset we draw one batch of size 128, and for the wild dataset $\mathbb{P}_{\text{wild}}$ we draw another batch of size 128 where each example is drawn from $\mathbb{P}_{\text{out}}^{\text{covariate}}$ with probability π_c, from $\mathbb{P}_{\text{out}}^{\text{semantic}}$ with probability π_s and from $\mathbb{P}_{\text{in}}$ with probability $1 - \pi_c - \pi_s$. For evaluation, we use the original test split of the ID data. Details of data split for OOD datasets are described in Appendix C. To evaluate each method including baselines, we use the collection of metrics defined in Section 2. The threshold for an energy-based OOD detector is selected based on the ID test set when 95% of ID test data points are declared as ID.

Training details. For CIFAR experiments and methods, we use the Wide ResNet (Zagoruyko & Komodakis, 2016) architecture with 40 layers and widen factor of 2. The model is optimized using stochastic gradient descent with Nesterov momentum (Duchi et al., 2011). We set the weight decay as 0.0005, and momentum as 0.09. We initialize the model with a pre-trained network on CIFAR-10, and then trained for 100 epochs using our method. The initial learning rate is 0.0001 and decays by a factor of 2 at epochs 50, 75, and 90. Following the previous literature (Katz-Samuels et al., 2022), we set $\alpha = 0.05$, and set τ to be twice the loss of the pre-trained model. For all the experiments, we use a batch size of 128 and a dropout rate of 0.3. Our framework was implemented with PyTorch 1.8.1. All training is performed using NVIDIA GeForce RTX 2080 Ti. See Appendix B for additional experimental details, including our validation strategy for selecting η.

4.2. Results and Discussion

Effect of margin η. Since the energy margin is central to our learning framework, we first aim to understand how η impacts performance. In Table 2, we perform an ablation by varying the margin $\eta \in \{0, -0.1, -0.5, -1, -2, -10, -20, -50\}$. Since energy should be negative for ID data, more negative values of η translate to a stronger margin constraint on ID points. As the margin changes from $\eta = 0$ to $\eta = -10$, a salient observation is that our method can substantially improve the OOD accuracy from 52.76% to 84.69%—a direct **31.93**% improvement in accuracy. At the same time, the ID accuracy remains comparable across different η, suggesting that one can indeed leverage the wild data to gain an improvement in OOD generalization without sacrificing ID classification accuracy. Furthermore, we observe that a tradeoff may exist between OOD generalization and OOD detection performance: the optimal OOD accuracy is achieved under a large margin, which can slightly degrade the OOD detection performance when compared to $\eta = 0$. Despite the tradeoff, we will show that SCONE still outperforms competing OOD detection methods (*c.f.* Table 3).

Table 2. Experimental results on CIFAR with different margin η. We train on CIFAR-10 as ID, using wild data with $\pi_c = 0.5$ (CIFAR-10-C) and $\pi_s = 0.1$ (SVHN).

margin	OOD Acc.↑	ID Acc.↑	FPR↓	AUROC↑
No margin	52.76	94.86	2.11	99.52
$\eta = -0.1$	53.24	94.87	2.16	99.52
$\eta = -0.5$	54.22	94.85	2.31	99.49
$\eta = -1$	55.55	94.88	2.56	99.45
$\eta = -2$	58.47	95.00	3.19	99.35
$\eta = -10$	84.69	94.65	10.86	97.84
$\eta = -20$	84.57	94.81	19.04	96.29
$\eta = -50$	84.56	94.83	19.24	96.25

SCONE achieves strong performance. We present the main results in Table 3, where SCONE establishes *overall strong performance in both OOD generalization and OOD detection*. In particular, we consider two broad categories of methods that are developed for either OOD detection or OOD generalization, and thus are expected to excel in only one of these two tasks. In contrast, our method targets both tasks simultaneously. Details of baseline implementation are in Appendix B.

We highlight a few observations: **(1)** SCONE outperforms competitive post hoc OOD detection methods, including `MSP` (Hendrycks & Gimpel, 2017), `ODIN` (Liang et al., 2018), `Energy` (Liu et al., 2020), `Mahalanobis` (Lee et al., 2018), `ViM` (Wang et al., 2022a), and the latest baseline `KNN` (Sun et al., 2022) — all of which use the same model trained with the CE loss on $\mathbb{P}_{\text{in}}$ (and hence they display the same OOD accuracy on CIFAR-10-C). As an example of our method's improved performance, when using SVHN as $\mathbb{P}_{\text{out}}^{\text{semantic}}$ our method yields an FPR95 of 10.86% (lower is better), which outperforms the best baseline performance of 12.89%. At the same time, the OOD generalization performance is significantly improved from 75.05% (baseline CE model) to 84.69% (ours). **(2)** Our method also outperforms common OOD generalization baselines, including `IRM` (Arjovsky et al., 2019), `Mixup` (Zhang et al., 2018), and `VREx` (Krueger et al., 2021). While these methods display stronger OOD generalization performance than the `ERM` (or `CE`) baseline, they underperform ours both in terms of OOD generalization and OOD detection.[2] **(3)** Lastly, we also compare to OOD detection methods utilizing $\mathbb{P}_{\text{wild}}$,

[2]Same as ours, we use the energy score in test time to perform OOD detection.

Table 3. Main results: comparison with competitive OOD generalization and OOD detection methods on CIFAR-10. We run our method 3 times and report the average and std. For experiments using $\mathbb{P}_{\text{wild}}$, we set $\pi_s = 0.5$, $\pi_c = 0.1$. For each semantic OOD dataset, we create corresponding wild mixture distribution $\mathbb{P}_{\text{wild}} := (1 - \pi_s - \pi_c)\mathbb{P}_{\text{in}} + \pi_s\mathbb{P}_{\text{out}}^{\text{semantic}} + \pi_c\mathbb{P}_{\text{out}}^{\text{covariate}}$ for training and evaluating on the corresponding test dataset. $\pm x$ denotes the standard error, rounded to the first decimal point. Results for LSUN-R and Texture datasets are in Appendix D. (*Since all the OOD detection methods use the same model trained with the CE loss on $\mathbb{P}_{\text{in}}$, they display the same ID and OOD accuracy on CIFAR-10-C.)

Method	SVHN $\mathbb{P}_{\text{out}}^{\text{semantic}}$, CIFAR-10-C $\mathbb{P}_{\text{out}}^{\text{covariate}}$				LSUN-C $\mathbb{P}_{\text{out}}^{\text{semantic}}$, CIFAR-10-C $\mathbb{P}_{\text{out}}^{\text{covariate}}$				Places365 $\mathbb{P}_{\text{out}}^{\text{semantic}}$, CIFAR-10-C $\mathbb{P}_{\text{out}}^{\text{covariate}}$			
	OOD Acc.↑	**ID Acc.↑**	**FPR↓**	**AUROC↑**	**OOD Acc.↑**	**ID Acc.↑**	**FPR↓**	**AUROC↑**	**OOD Acc.↑**	**ID Acc.↑**	**FPR↓**	**AUROC↑**
OOD detection												
MSP	75.05	94.84	48.49	91.89	75.05	94.84	30.80	95.65	75.05	94.84	57.40	84.49
ODIN	75.05	94.84	33.35	91.96	75.05	94.84	15.52	97.04	75.05	94.84	57.40	84.49
Energy	75.05	94.84	35.59	90.96	75.05	94.84	8.26	98.35	75.05	94.84	40.14	89.89
Mahalanobis	75.05	94.84	12.89	97.62	75.05	94.84	39.22	94.15	75.05	94.84	68.57	84.61
ViM	75.05	94.84	21.95	95.48	75.05	94.84	5.90	98.82	75.05	94.84	21.95	95.48
KNN	75.05	94.84	28.92	95.71	75.05	94.84	28.08	95.33	75.05	94.84	42.67	91.07
OOD generalization												
ERM	75.05	94.84	35.59	90.96	75.05	94.84	8.26	98.35	75.05	94.84	40.14	89.89
Mixup	79.17	93.30	97.33	18.78	79.17	93.30	52.10	76.66	79.17	93.30	58.24	75.70
IRM	77.92	90.85	63.65	90.70	77.92	90.85	36.67	94.22	77.92	90.85	53.79	88.15
VREx	76.90	91.35	55.92	91.22	76.90	91.35	51.50	91.56	76.90	91.35	56.13	87.45
Learning w. $\mathbb{P}_{wild}$												
OE	37.61	94.68	0.84	99.80	41.37	93.99	3.07	99.26	35.98	94.75	27.02	94.57
Energy (w. outlier)	20.74	90.22	0.86	99.81	32.55	92.97	2.33	99.93	19.86	90.55	23.89	93.60
Woods	52.76	94.86	2.11	99.52	76.90	95.02	1.80	99.56	54.58	94.88	30.48	93.28
Scone (ours)	$84.69_{\pm 0.1}$	$94.65_{\pm 0.0}$	$10.86_{\pm 0.7}$	$97.84_{\pm 0.1}$	$84.58_{\pm 0.7}$	$93.73_{\pm 0.4}$	$10.23_{\pm 1.1}$	$98.02_{\pm 0.2}$	$85.21_{\pm 0.1}$	$94.59_{\pm 0.0}$	$37.56_{\pm 0.2}$	$90.90_{\pm 0.1}$

Table 4. Ablations on mixing ratios π_c. We train on CIFAR-10 as ID, using CIFAR-10-C for $\mathbb{P}_{\text{ood}}^{\text{covariate}}$ and SVHN for $\mathbb{P}_{\text{ood}}^{\text{semantic}}$ (with fixed $\pi_s = 0.1$). For our method, $\eta = -10$, which is chosen based on the validation procedure in Appendix B.

π_c	Method	OOD Acc.↑	ID Acc.↑	FPR↓	AUROC↑
0.0	WOODS	76.44	94.92	1.18	99.78
0.0	SCONE	76.52	94.93	1.60	99.71
0.1	WOODS	73.65	94.92	1.47	99.72
0.1	SCONE	78.27	94.88	10.61	97.91
0.2	WOODS	65.55	95.02	2.19	99.57
0.2	SCONE	80.70	94.93	9.97	98.04
0.5	WOODS	52.76	94.86	2.11	99.52
0.5	SCONE	84.69	94.65	10.86	97.84
0.9	WOODS	52.84	94.81	1.80	99.57
0.9	SCONE	86.18	94.64	13.68	97.41

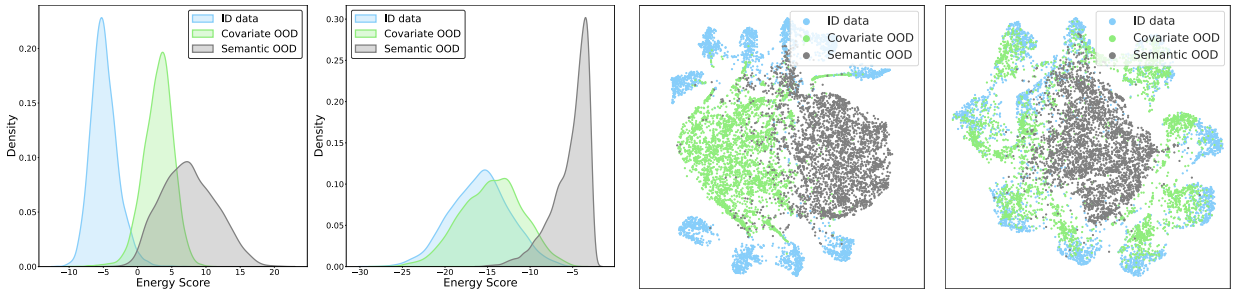
including OE (Hendrycks et al., 2018), energy-regularized learning (Liu et al., 2020), and WOODS (Katz-Samuels et al., 2022), which is the latest such method to be developed. These methods are among the strongest OOD detection methods, yet display a significantly worsened OOD generalization performance. The main reason is that they make assumptions on $\mathbb{P}_{\text{wild}}$ without considering covariate OOD data.

Effect of different mixing ratios. In Table 4, we ablate the effect of π_c, which modulates the fraction of covariate OOD data in the mixture distribution $\mathbb{P}_{\text{wild}}$. For all settings, we contrast the performance without vs. with margin. The margin in each π_c setting is validated using the strategy in Appendix B. Here we consistently use $\pi_s = 0.1$, which reflects the practical scenario that the majority of test data may remain in the known classes. We primarily focus on evaluations where $\pi_c \neq 0$, since our problem setting uniquely introduces the covariate shift in the wild distribution. However, for completeness, we also include results when $\pi_c = 0$. We highlight a few interesting observations: **(1)** without any enforced margin, the OOD generalization performance for WOODS (Katz-Samuels et al., 2022) generally degrades with increasing π_c. This is likely due to the fact that a larger π_c translates into more severe covariate shifts. For example, when $\pi_c = 0.9$, the OOD classification accuracy decreases to 52.84%. **(2)** Our method is overall more robust under large π_c settings than the WOODS baseline. For instance, in a challenging case with $\pi_c = 0.9$, SCONE outperforms WOODS by **33.34%**. Overall, these results demonstrate the benefits of SCONE for both OOD generalization and detection.

4.3. Qualitative Insights

Visualization of OOD score distributions. We visualize the energy score distribution in Figure 3 (a) and (b), for WOODS vs. our method, respectively. There are two salient observations: first, the energy scores for ID data indeed shift from -5 (without margin) towards more negative values (e.g., -16), suggesting the efficacy of our margin-based optimization. Moreover, the energy score distributions between $\mathbb{P}_{\text{in}}$ and $\mathbb{P}_{\text{out}}^{\text{covariate}}$ becomes more aligned than in WOODS. This can be attributed to the aligned feature representation, which we verify next.

Visualization of feature embeddings. Figure 3 shows t-SNE visualizations (Van der Maaten & Hinton, 2008) of the



(a) Energy scores of WOODS (b) Energy scores of Ours (c) T-SNE of WOODS (d) T-SNE of Ours

Figure 3. (a)-(b) Energy score distributions for WOODS vs. our method. Different colors represent the different types of test data: CIFAR-10 as $\mathbb{P}_{\text{in}}$ (blue), CIFAR-10-C as $\mathbb{P}_{\text{out}}^{\text{covariate}}$ (green), and SVHN as $\mathbb{P}_{\text{out}}^{\text{semantic}}$ (gray). (c)-(d): T-SNE visualization of the image embeddings using WOODS vs. our method.

Table 5. Results on ImageNet-100. We use ImageNet-100 as ID, ImageNet-100-C for $\mathbb{P}_{\text{ood}}^{\text{covariate}}$ and iNaturalist for $\mathbb{P}_{\text{ood}}^{\text{semantic}}$.

Method	**OOD Accuracy (ImageNet-100-C)** (OOD generalization) ↑	**ID Accuracy (ImageNet-100)** (ID generalization) ↑	**FPR95** (OOD detection) ↓	**AUROC** (OOD detection) ↑
WOODS (Katz-Samuels et al., 2022)	44.46	86.49	10.50	98.22
SCONE (ours)	65.34	87.64	27.13	95.66

penultimate-layer feature embedding, for WOODS (c) vs. our method (d). The embeddings are extracted from the test split. The blue points denote the test ID set (CIFAR-10), the green points are the test samples from CIFAR-10-C, and the gray points are from SVHN. This visualization suggests that embeddings of CIFAR and CIFAR-C become more aligned with our margin-based optimization, which arguably leads to improved OOD generalization performance. This observation corroborates our theoretical insight in Section 3.3.

4.4. Experiments on ImageNet

In this section, we provide additional large-scale results on the ImageNet benchmark. We use ImageNet-100 as the in-distribution data ($\mathbb{P}_{\text{in}}$), with labels provided in Appendix C. For the covariate-shifted OOD data ($\mathbb{P}_{\text{out}}^{\text{covariate}}$), we use ImageNet-100-C with Gaussian noise in the experiment. For the semantic-shifted OOD data, we use the high-resolution natural images from iNaturalist (Van Horn et al., 2018), with the same subset as MOS (Huang & Li, 2021). The wild data $\mathbb{P}_{\text{in}}$ is a mixture of ID data, covariate-shifted data ($\pi_c = 0.5$), and semantic shifted data ($\pi_s = 0.1$). We fine-tune ResNet-34 (He et al., 2016) (pre-trained on ImageNet) for 100 epochs, with an initial learning rate of 0.01 and a batch size of 64. Results in Table 5 suggest that our method can improve both ID and OOD accuracy compared to WOODS (the most competitive baseline).

5. Related Works

Out-of-distribution detection is of vital importance for machine learning models deployed in the open world. Recent advances in OOD detection can be broadly categorized into post hoc and regularization-based methods. In particular, post hoc methods (Hendrycks & Gimpel, 2017; Liang et al., 2018; Lee et al., 2018; Liu et al., 2020; Huang et al., 2021; Sun et al., 2021; 2022) focus on deriving test-time OOD scoring functions for a pre-trained classifier. Our proposed work is closer to another line of work (Bevandić et al., 2018; Hendrycks et al., 2018; Malinin & Gales, 2018; Liu et al., 2020; Du et al., 2021; Ming et al., 2022), which addresses the OOD detection problem by training-time regularization. For example, models are encouraged to give predictions with lower confidence (Hendrycks et al., 2018) or higher energies (Liu et al., 2020). These methods require access to a clean OOD dataset for training, which can be restrictive. To circumvent this, a recent work WOODS (Katz-Samuels et al., 2022) first explored using wild mixture data consisting of the ID and semantically shifted OOD data $\mathbb{P}_{\text{wild}} := (1-\pi)\mathbb{P}_{\text{in}} + \pi\mathbb{P}_{\text{out}}^{\text{semantic}}$. The crucial difference between our work and WOODS is whether the wild mixture data contains covariate-shifted data, which introduces new challenges not considered in prior work. As we show in this work, our formulation uniquely enables both OOD generalization and OOD detection, in one coherent framework.

Out-of-distribution generalization is a fundamental problem in machine learning, which aims to generalize to covariate-shifted data without any sample from the target domain (Muandet et al., 2013; Arjovsky et al., 2019; Bahng et al., 2020; Wang et al., 2022b; Xie et al., 2021). OOD generalization is more challenging compared to the classic domain adaptation problem (Daume III & Marcu, 2006; Blitzer et al., 2007; Ben-David et al., 2010; Ganin & Lem-

pitsky, 2015; Tzeng et al., 2017; Redko et al., 2019; Kang et al., 2019; Kumar et al., 2020; Wang et al., 2022c), which assumes access to labeled samples from the target domain.

To highlight a few works, IRM (Arjovsky et al., 2019) and its variants (Krueger et al., 2021; Ahuja et al., 2020) aim to find invariant representation from different training environments via an invariant risk regularizer. GroupDRO (Sagawa et al., 2019) and Probabilistic Group DRO (Ghosal & Li, 2023) minimize the worst-case training loss over a set of groups. Zhou et al. (2020) propose generating outlier samples of a novel domain, which are then used for improving the generalization of the classifier. Besides algorithm innovation, benchmark efforts have also been pursued by DomainBed (Gulrajani & Lopez-Paz, 2020) and OoD-Bench (Ye et al., 2022), which facilitates evaluation on OOD generalization. OS-SDG (Zhu & Li) relies on one source domain with labels to train the model, while our framework can exploit unlabeled wild data naturally arising in the wild, which is a mixed composition of three data distributions. Different from previous literature, we focus on improving OoD robustness in classifiers by learning from the wild mixture data and building an OOD detector at the same time. To the best of our knowledge, we are the first work that leverages wild data for both OOD generalization and OOD detection purposes. Our framework also uniquely allows leveraging covariate-shifted data freely arising in the wild, without requiring any labeling.

Universal domain adaptation aims to leverage labeled data from a related domain (source domain) and improve the model performance for the target domain, where there exists category gap for label sets between the source and target domains (You et al., 2019). Several works have been proposed to address this problem (Saito et al., 2020; Fu et al., 2020; Li et al., 2021; Chen et al., 2022; Kundu et al., 2022; Chang et al., 2022; Garg et al., 2022). UAN (You et al., 2019) presents a universal adaptation network that exploits both the domain similarity and prediction uncertainty of each sample for promoting common-class adaptation. DANCE (Saito et al., 2020) proposes a domain adaptative neighborhood clustering technique for category shift-agnostic adaptation via entropy optimization. The work in Chang et al. (2022) proposes a unified optimal transport-based framework to encourage both global cluster discrimination and local consistency of samples. Different from prior works, we leverage both labeled in-distribution data and unlabeled wild data when training our model. Such unlabeled wild data naturally arise in real-world environments and have not been considered in prior literature.

Positive-Unlabeled (PU) learning is a classic machine learning problem, which aims to learn classifiers from positive and unlabeled data (Letouzey et al., 2000). Multiple prior works have been proposed for discussing PU learning (Hsieh et al., 2015; Zhao et al., 2022; Acharya et al., 2022; Chapel et al., 2020; Xu & Denil, 2021). The work in (Niu et al., 2016) proposes a theoretical comparison of positive-unlabeled learning against positive-negative learning based on the upper bounds of estimation errors. Du Plessis et al. (2015) presents a convex formulation for PU learning by using different loss functions for positive and unlabeled samples. Margin-based PU learning (Gong et al., 2018) introduces a provable positive margin-based PU learning algorithm for classification under the truncated linear distributions. There are two key differences between ours and PU learning: (1) PU learning only considers the task of distinguishing $\mathbb{P}_{\text{out}}$ (anomalous) and $\mathbb{P}_{\text{in}}$ (normal), not the task of doing classification simultaneously. We consider OOD detection which additionally requires learning a classifier for the distribution $\mathbb{P}_{\mathcal{X}\mathcal{Y}}$. (2) PU learning does not consider the generalization aspect under covariate-shifted OOD data, whereas our framework handles it in addition to semantic-shifted OOD data.

6. Conclusion

In this study, we propose a novel framework SCONE to jointly tackle the OOD generalization and OOD detection problems by leveraging wild data—a mixture of ID, covariate OOD, and semantic OOD data. Our framework offers practical advantages since the wild data is freely collectible in abundance, does not require any human annotation, and importantly, captures the environmental test-time OOD distributions under both covariate and semantic shifts. We make use of such unlabeled wild data to train a binary OOD detector, and at the same time, enhance the generalization ability of the ID classifier. We provide new theoretical and empirical insights on the importance of enforcing a sufficient margin between the OOD decision boundary and ID data. Extensive experiments show that our framework can effectively improve both OOD generalization and detection performance. We hope our framework will inspire both OOD generalization and OOD detection communities to tackle data shift problems synergistically.

Acknowledgement

The work is supported in part by the AFOSR Young Investigator Award under No. FA9550-23-1-0184; Philanthropic Fund from SFF; Wisconsin Alumni Research Foundation (WARF); and faculty research awards/gifts from Google, Meta, and Amazon. Any opinions, findings, conclusions, or recommendations expressed in this material are those of the authors and do not necessarily reflect the views, policies, or endorsements either expressed or implied, of the sponsors. The authors would also like to thank ICML reviewers for their helpful suggestions and feedback.

References

Acharya, A., Sanghavi, S., Jing, L., Bhushanam, B., Choudhary, D., Rabbat, M., and Dhillon, I. Positive unlabeled contrastive learning. *arXiv preprint arXiv:2206.01206*, 2022.

Ahuja, K., Shanmugam, K., Varshney, K., and Dhurandhar, A. Invariant risk minimization games. In *International Conference on Machine Learning*, pp. 145–155. PMLR, 2020.

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

Bahng, H., Chun, S., Yun, S., Choo, J., and Oh, S. J. Learning de-biased representations with biased representations. In *International Conference on Machine Learning*, pp. 528–539. PMLR, 2020.

Ben-David, S., Blitzer, J., Crammer, K., Kulesza, A., Pereira, F., and Vaughan, J. W. A theory of learning from different domains. *Machine learning*, 79(1):151–175, 2010.

Bevandić, P., Krešo, I., Oršić, M., and Šegvić, S. Discriminative out-of-distribution detection for semantic segmentation. *arXiv preprint arXiv:1808.07703*, 2018.

Blanchard, G., Deshmukh, A. A., Dogan, Ü., Lee, G., and Scott, C. Domain generalization by marginal transfer learning. *The Journal of Machine Learning Research*, 22 (1):46–100, 2021.

Blitzer, J., Crammer, K., Kulesza, A., Pereira, F., and Wortman, J. Learning bounds for domain adaptation. *Advances in neural information processing systems*, 20, 2007.

Chang, W., Shi, Y., Tuan, H., and Wang, J. Unified optimal transport framework for universal domain adaptation. *Advances in Neural Information Processing Systems*, 35: 29512–29524, 2022.

Chapel, L., Alaya, M. Z., and Gasso, G. Partial optimal tranport with applications on positive-unlabeled learning. *Advances in Neural Information Processing Systems*, 33: 2903–2913, 2020.

Chen, L., Lou, Y., He, J., Bai, T., and Deng, M. Geometric anchor correspondence mining with uncertainty modeling for universal domain adaptation. In *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 16134–16143, 2022.

Cimpoi, M., Maji, S., Kokkinos, I., Mohamed, S., and Vedaldi, A. Describing textures in the wild. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 3606–3613, 2014.

Daume III, H. and Marcu, D. Domain adaptation for statistical classifiers. *Journal of artificial Intelligence research*, 26:101–126, 2006.

Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K., and Fei-Fei, L. Imagenet: A large-scale hierarchical image database. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pp. 248–255. Ieee, 2009.

Du, X., Wang, Z., Cai, M., and Li, Y. Vos: Learning what you don't know by virtual outlier synthesis. In *International Conference on Learning Representations*, 2021.

Du Plessis, M., Niu, G., and Sugiyama, M. Convex formulation for learning from positive and unlabeled data. In *International Conference on Machine Learning*, pp. 1386–1394. PMLR, 2015.

Duchi, J., Hazan, E., and Singer, Y. Adaptive subgradient methods for online learning and stochastic optimization. *Journal of Machine Learning Research*, 12(7), 2011.

Fu, B., Cao, Z., Long, M., and Wang, J. Learning to detect open classes for universal domain adaptation. In *European Conference on Computer Vision*, pp. 567–583, 2020.

Ganin, Y. and Lempitsky, V. Unsupervised domain adaptation by backpropagation. In *International conference on machine learning*, pp. 1180–1189. PMLR, 2015.

Ganin, Y., Ustinova, E., Ajakan, H., Germain, P., Larochelle, H., Laviolette, F., Marchand, M., and Lempitsky, V. Domain-adversarial training of neural networks. *The Journal of Machine Learning Research*, 17(1):2096–2030, 2016.

Garg, S., Balakrishnan, S., and Lipton, Z. Domain adaptation under open set label shift. *Advances in Neural Information Processing Systems*, 35:22531–22546, 2022.

Ghosal, S. S. and Li, Y. Distributionally robust optimization with probabilistic group. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 2023.

Gong, T., Wang, G., Ye, J., Xu, Z., and Lin, M. Margin based pu learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 32, 2018.

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

He, K., Zhang, X., Ren, S., and Sun, J. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 770–778, 2016.

Hendrycks, D. and Dietterich, T. G. Benchmarking neural network robustness to common corruptions and surface variations. *arXiv preprint arXiv:1807.01697*, 2018.

Hendrycks, D. and Gimpel, K. A baseline for detecting misclassified and out-of-distribution examples in neural networks. *International Conference on Learning Representations*, 2017.

Hendrycks, D., Mazeika, M., and Dietterich, T. Deep anomaly detection with outlier exposure. In *International Conference on Learning Representations*, 2018.

Hestenes, M. R. Multiplier and gradient methods. *Journal of Optimization Theory and Applications*, 4(5):303–320, 1969.

Hsieh, C.-J., Natarajan, N., and Dhillon, I. Pu learning for matrix completion. In *International Conference on Machine Learning*, pp. 2445–2453. PMLR, 2015.

Huang, R. and Li, Y. Mos: Towards scaling out-of-distribution detection for large semantic space. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 8710–8719, 2021.

Huang, R., Geng, A., and Li, Y. On the importance of gradients for detecting distributional shifts in the wild. *Advances in Neural Information Processing Systems*, 34: 677–689, 2021.

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

Kang, G., Jiang, L., Yang, Y., and Hauptmann, A. G. Contrastive adaptation network for unsupervised domain adaptation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 4893–4902, 2019.

Katz-Samuels, J., Nakhleh, J. B., Nowak, R., and Li, Y. Training ood detectors in their natural habitats. In *International Conference on Machine Learning*. PMLR, 2022.

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.

Krizhevsky, A., Hinton, G., et al. Learning multiple layers of features from tiny images. *Technical report, University of Toronto*, 2009.

Krueger, D., Caballero, E., Jacobsen, J.-H., Zhang, A., Binas, J., Zhang, D., Le Priol, R., and Courville, A. Out-of-distribution generalization via risk extrapolation (rex). In *International Conference on Machine Learning*, pp. 5815–5826. PMLR, 2021.

Kumar, A., Ma, T., and Liang, P. Understanding self-training for gradual domain adaptation. In *International Conference on Machine Learning*, pp. 5468–5479. PMLR, 2020.

Kundu, J. N., Bhambri, S., Kulkarni, A. R., Sarkar, H., Jampani, V., et al. Subsidiary prototype alignment for universal domain adaptation. *Advances in Neural Information Processing Systems*, 35:29649–29662, 2022.

Langley, P. Crafting papers on machine learning. In Langley, P. (ed.), *Proceedings of the 17th International Conference on Machine Learning (ICML 2000)*, pp. 1207–1216, Stanford, CA, 2000. Morgan Kaufmann.

LeCun, Y. The mnist database of handwritten digits. *http://yann. lecun. com/exdb/mnist/*, 1998.

Lee, K., Lee, K., Lee, H., and Shin, J. A simple unified framework for detecting out-of-distribution samples and adversarial attacks. *Advances in neural information processing systems*, 31, 2018.

Letouzey, F., Denis, F., and Gilleron, R. Learning from positive and unlabeled examples. In *International Conference on Algorithmic Learning Theory*, pp. 71–85. Springer, 2000.

Li, D., Yang, Y., Song, Y.-Z., and Hospedales, T. M. Deeper, broader and artier domain generalization. In *IEEE International Conference on Computer Vision*, pp. 5542–5550, 2017.

Li, D., Yang, Y., Song, Y.-Z., and Hospedales, T. Learning to generalize: Meta-learning for domain generalization. In *Proceedings of the AAAI conference on artificial intelligence*, volume 32, 2018a.

Li, G., Kang, G., Zhu, Y., Wei, Y., and Yang, Y. Domain consensus clustering for universal domain adaptation. In *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 9757–9766, 2021.

Li, H., Pan, S. J., Wang, S., and Kot, A. C. Domain generalization with adversarial feature learning. In *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 5400–5409, 2018b.

Li, Y., Tian, X., Gong, M., Liu, Y., Liu, T., Zhang, K., and Tao, D. Deep domain generalization via conditional invariant adversarial networks. In *European Conference on Computer Vision*, pp. 624–639, 2018c.

Liang, S., Li, Y., and Srikant, R. Enhancing the reliability of out-of-distribution image detection in neural networks. In *International Conference on Learning Representations*, 2018.

Liu, W., Wang, X., Owens, J., and Li, Y. Energy-based out-of-distribution detection. *Advances in Neural Information Processing Systems*, 2020.

Malinin, A. and Gales, M. Predictive uncertainty estimation via prior networks. *arXiv preprint arXiv:1802.10501*, 2018.

Ming, Y., Fan, Y., and Li, Y. Poem: Out-of-distribution detection with posterior sampling. In *International Conference on Machine Learning (ICML)*. PMLR, 2022.

Mu, N. and Gilmer, J. Mnist-c: A robustness benchmark for computer vision. *International Conference on Machine Learning Workshop*, 2019.

Muandet, K., Balduzzi, D., and Schölkopf, B. Domain generalization via invariant feature representation. In *International Conference on Machine Learning*, pp. 10–18. PMLR, 2013.

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

Netzer, Y., Wang, T., Coates, A., Bissacco, A., Wu, B., and Ng, A. Y. Reading digits in natural images with unsupervised feature learning. *Neural Information Processing Systems Workshops*, 2011.

Niu, G., du Plessis, M. C., Sakai, T., Ma, Y., and Sugiyama, M. Theoretical comparisons of positive-unlabeled learning against positive-negative learning. *Advances in Neural Information Processing Systems*, 29, 2016.

Parhi, R. and Nowak, R. D. What kinds of functions do deep neural networks learn? insights from variational spline theory. *SIAM Journal on Mathematics of Data Science*, 4(2):464–489, 2022. doi: 10.1137/21M1418642. URL https://doi.org/10.1137/21M1418642.

Redko, I., Morvant, E., Habrard, A., Sebban, M., and Bennani, Y. *Advances in domain adaptation theory*. Elsevier, 2019.

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

Saito, K., Kim, D., Sclaroff, S., and Saenko, K. Universal domain adaptation through self-supervision. *Advances in Neural Information Processing Systems*, 33:16282–16292, 2020.

Sun, B. and Saenko, K. Deep coral: Correlation alignment for deep domain adaptation. In *European Conference on Computer Vision*, pp. 443–450, 2016.

Sun, Y., Guo, C., and Li, Y. React: Out-of-distribution detection with rectified activations. In *Advances in Neural Information Processing Systems*, 2021.

Sun, Y., Ming, Y., Zhu, X., and Li, Y. Out-of-distribution detection with deep nearest neighbors. *International Conference on Machine Learning*, 2022.

Tzeng, E., Hoffman, J., Saenko, K., and Darrell, T. Adversarial discriminative domain adaptation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 7167–7176, 2017.

Van der Maaten, L. and Hinton, G. Visualizing data using t-sne. *Journal of machine learning research*, 9(11), 2008.

Van Horn, G., Mac Aodha, O., Song, Y., Cui, Y., Sun, C., Shepard, A., Adam, H., Perona, P., and Belongie, S. The inaturalist species classification and detection dataset. In *Proceedings of the IEEE Conference on Computer Vision and Pattern recognition*, pp. 8769–8778, 2018.

Vapnik, V. N. An overview of statistical learning theory. *IEEE Transactions on Neural Networks*, 10(5):988–999, 1999.

Wang, H., Li, Z., Feng, L., and Zhang, W. Vim: Out-of-distribution with virtual-logit matching. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 4921–4930, 2022a.

Wang, J., Lan, C., Liu, C., Ouyang, Y., Qin, T., Lu, W., Chen, Y., Zeng, W., and Yu, P. Generalizing to unseen domains: A survey on domain generalization. *IEEE Transactions on Knowledge and Data Engineering*, 2022b.

Wang, R., Chaudhari, P., and Davatzikos, C. Embracing the disharmony in medical imaging: A simple and effective framework for domain adaptation. *Medical Image Analysis*, 76:102309, 2022c.

Wang, Y., Li, H., and Kot, A. C. Heterogeneous domain generalization via domain mixup. In *IEEE International Conference on Acoustics, Speech and Signal Processing*, pp. 3622–3626, 2020.

Xiao, H., Rasul, K., and Vollgraf, R. Fashion-mnist: a novel image dataset for benchmarking machine learning algorithms. *arXiv preprint arXiv:1708.07747*, 2017.

Xie, S. M., Kumar, A., Jones, R., Khani, F., Ma, T., and Liang, P. In-n-out: Pre-training and self-training using auxiliary information for out-of-distribution robustness. In *International Conference on Learning Representations*, 2021.

Xu, D. and Denil, M. Positive-unlabeled reward learning. In *Conference on Robot Learning*, pp. 205–219. PMLR, 2021.

Xu, M., Zhang, J., Ni, B., Li, T., Wang, C., Tian, Q., and Zhang, W. Adversarial domain adaptation with domain mixup. In *AAAI Conference on Artificial Intelligence*, volume 34, pp. 6502–6509, 2020.

Yang, J., Zhou, K., Li, Y., and Liu, Z. Generalized out-of-distribution detection: A survey. *arXiv preprint arXiv:2110.11334*, 2021.

Ye, N., Li, K., Bai, H., Yu, R., Hong, L., Zhou, F., Li, Z., and Zhu, J. Ood-bench: Quantifying and understanding two dimensions of out-of-distribution generalization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 7947–7958, 2022.

You, K., Long, M., Cao, Z., Wang, J., and Jordan, M. I. Universal domain adaptation. In *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 2720–2729, 2019.

Yu, F., Seff, A., Zhang, Y., Song, S., Funkhouser, T., and Xiao, J. Lsun: Construction of a large-scale image dataset using deep learning with humans in the loop. *arXiv preprint arXiv:1506.03365*, 2015.

Zagoruyko, S. and Komodakis, N. Wide residual networks. *arXiv preprint arXiv:1605.07146*, 2016.

Zhang, H., Cisse, M., Dauphin, Y. N., and Lopez-Paz, D. mixup: Beyond empirical risk minimization. In *International Conference on Learning Representations*, 2018.

Zhang, M. M., Marklund, H., Dhawan, N., Gupta, A., Levine, S., and Finn, C. Adaptive risk minimization: Learning to adapt to domain shift. In *Advances in Neural Information Processing Systems*, 2021.

Zhao, Y., Xu, Q., Jiang, Y., Wen, P., and Huang, Q. Dist-pu: Positive-unlabeled learning from a label distribution perspective. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 14461–14470, 2022.

Zhou, B., Lapedriza, A., Khosla, A., Oliva, A., and Torralba, A. Places: A 10 million image database for scene recognition. *IEEE transactions on pattern analysis and machine intelligence*, 40(6):1452–1464, 2017.

Zhou, K., Yang, Y., Hospedales, T., and Xiang, T. Learning to generate novel domains for domain generalization. In *European conference on computer vision*, pp. 561–578. Springer, 2020.

Zhu, R. and Li, S. Crossmatch: Cross-classifier consistency regularization for open-set single domain generalization. In *International Conference on Learning Representations*.

A. Proof of Proposition 3.1

In this setting, the energy function becomes $E_\theta(\phi(\mathbf{x})) = -\log\left[e^{\frac{1}{2}\overline{f}_\theta(\phi(\mathbf{x}))} + e^{-\frac{1}{2}\overline{f}_\theta(\phi(\mathbf{x}))}\right]$. If $E_\theta(\phi(\mathbf{x})) < \eta$, then due to the symmetry of $e^u + e^{-u}$ we have $\left|\overline{f}_\theta(\phi(\mathbf{x}))\right| > -2\eta - 2\log 2$. Therefore, setting $\eta < 0$ effectively lower bounds the distance of ID points from the classification decision boundary in this setting.

Consider the value of $\overline{f}_\theta(\phi(\mathbf{x}_c))$ for a covariate shifted point $\mathbf{x}_c$ with label $y = 1$. By assumption, there exists an ID point $\mathbf{x}$, also with label $y = 1$, satisfying $\|\phi(\mathbf{x}_c) - \phi(\mathbf{x})\|_2 < \delta$. We have

$$\begin{aligned}
\overline{f}_\theta(\phi(\mathbf{x})) &= (\overline{f}_\theta(\phi(\mathbf{x})) - \overline{f}_\theta(\phi(\mathbf{x}_c))) + \overline{f}_\theta(\phi(\mathbf{x}_c)) \\
&\leq \left|\overline{f}_\theta(\phi(\mathbf{x})) - \overline{f}_\theta(\phi(\mathbf{x}_c))\right| + \overline{f}_\theta(\phi(\mathbf{x}_c)) \\
&\leq L\left\|\phi(\mathbf{x}) - \phi(\mathbf{x}_c)\right\|_2 + \overline{f}_\theta(\phi(\mathbf{x}_c)) \\
&\leq L\delta + \overline{f}_\theta(\phi(\mathbf{x}_c)).
\end{aligned}$$

Since $\alpha = 0$, $\tau = 0$ and $y = 1$, $\overline{f}(\phi(\mathbf{x})) = \left|\overline{f}(\phi(\mathbf{x}))\right| > -2\eta - 2\log 2$, which combined with the above implies $\overline{f}(\phi(\mathbf{x}_c)) > -2\eta - 2\log 2 - L\delta$. A similar argument shows that if $y = 2$, $\overline{f}(\phi(\mathbf{x}_c)) < -(-2\eta - 2\log 2 - L\delta)$. Therefore, one can set $\eta < -\log 2 - \frac{1}{2}L\delta$ and ensure that $\mathbf{x}_c$ is classified correctly, with $\overline{f}(\phi(\mathbf{x}_c)) > 0$ when $y = 1$ and $\overline{f}(\phi(\mathbf{x}_c)) < 0$ when $y = 2$. We also immediately have an upper bound on $E_\theta(\mathbf{x}_c)$, since

$$E_\theta(\mathbf{x}_c) \leq -\frac{1}{2}\left|\overline{f}_\theta(\mathbf{x}_c)\right| \leq \eta + \log 2.$$

Under the setting $\eta < -\log 2 - \frac{1}{2}L\delta$, $E_\theta(\mathbf{x}_c) < -\frac{1}{2}L\delta < 0$, and so $g_\theta(\mathbf{x}_c) = \text{IN}$.

B. Experimental Details

Validation strategy for selecting η. Here we discuss how to choose the optimal margin parameter η from $\{0, -0.1, -0.5, -1, -2, -10, -20, -50\}$. A major challenge is that one may not have access to a clean validation set of either $\mathbb{P}_{\text{ood}}^{\text{covariate}}$ or $\mathbb{P}_{\text{ood}}^{\text{semantic}}$. More realistically, one may have a separate unlabeled set sampled from the wild mixed distribution $\mathbb{P}_{\text{wild}}$. We thus leverage this *mixed* dataset $\mathcal{D}_{\text{val}}$ for validation, and propose the following heuristic measurement that can help reliably select a good η:

$$\tilde{\text{out}}\% = \frac{\sum_{\tilde{\mathbf{x}}_i \in \mathcal{D}_{\text{val}}} \mathbb{1}\{g_\theta(\tilde{\mathbf{x}}_i) = \text{out}\}}{|\mathcal{D}_{\text{val}}|}.$$

This heuristic measures the fraction of samples in the validation set predicted as OUT by the OOD detector. We select η based on a "phase transition" under this measurement. We exemplify this in Table 6, based on our main experimental setting with CIFAR-10 as ID, CIFAR-10-C as Covariate-OOD, and SVHN as Semantic-OOD. The first phase (e.g., $\eta = 0, 0.1, 0.5, 1, 2$) corresponds to an OOD detector that classifies ID on one side and remainder samples (including covariate shifted ones) to be on the other side. As the margin enlarges further (e.g., $\eta = -10$), the OOD detector primarily identifies $\mathbb{P}_{\text{ood}}^{\text{semantic}}$ as OUT, which matches more closely with the desired behavior of OOD detector. This behavior translates into a drop in $\tilde{\text{out}}\%$. We use the η value corresponding to the drop as the selected margin parameter.

Table 6. Experimental results on CIFAR with different margin settings η. We train on CIFAR-10 as ID, using the same wild data with $\pi_c = 0.5$ (CIFAR-10-C) and $\pi_s = 0.1$ (SVHN).

margin	**OOD Acc.**↑	**ID Acc.**↑	**FPR**↓	**AUROC**↑	$\tilde{\text{out}}\%$
No margin	52.77	94.87	2.11	99.52	58.49
$\eta = -0.1$	53.24	94.87	2.16	99.52	58.32
$\eta = -0.5$	54.22	94.85	2.31	99.49	58.16
$\eta = -1$	55.55	94.88	2.56	99.45	57.53
$\eta = -2$	58.47	95.00	3.19	99.35	55.54
$\eta = -10$	84.69	94.65	10.86	97.84	17.30
$\eta = -20$	84.57	94.81	19.04	96.29	16.23
$\eta = -50$	84.56	94.83	19.24	96.25	16.20

Implementation details of baselines for OOD detection. We use Wide ResNet (Zagoruyko & Komodakis, 2016) with 40 layers and widen factor of 2. For evaluating the post hoc OOD detection baselines, we use the model trained with the CE loss on $\mathbb{P}_{\text{in}}$. We employ the same pre-trained model from the github: `https://github.com/wetliu/energy_ood`, which was trained on the complete CIFAR-10 dataset. Specifically, the model is trained using cross-entropy loss for 200 epochs. The learning rate is started at 0.1 and then decays by multiplier 0.1 at 100, 150, and 175 epochs. To facilitate easy comparison, the pre-trained model and baseline results are consistent with (Katz-Samuels et al., 2022) (courtesy of Table 4).

Implementation details of baselines for OOD generalization. For OOD generalization baselines, we use the same network architecture, Wide ResNet-40-2 (Zagoruyko & Komodakis, 2016), and train from scratch using respective losses. The baselines in Table 3 are trained on CIFAR-10 (Krizhevsky et al., 2009) $50,000$ labeled training examples. We follow the default hyperparameter setting as in original papers whenever applicable. For Mixup (Zhang et al., 2018), we follow the original paper and set hyperparameter $\alpha = 1$ to control the strength of interpolation between feature-target pairs. The λ for IRM (Arjovsky et al., 2019) baseline is set to 100 and the λ for VREx (Krueger et al., 2021) is set to 10 for the penalty weights. Following the original training configuration in WRN (Zagoruyko & Komodakis, 2016), all the baselines are trained for 200 epochs with batch size 128. We use the SGD optimizer with an initial learning rate of 0.1. The learning rate decays by a factor of 10 after 60, 120, and 180 epochs. Weight decay is set to 10^{-4}. All models are implemented in PyTorch 1.8.1. We evaluate the trained model on the CIFAR-10 test set (ID accuracy) and CIFAR-10-C (OOD accuracy).

C. Details of Datasets

We provide a detailed description of the datasets used in this work below:

MNIST (LeCun, 1998) is a large database of handwritten digits with 10 categories and is widely used in the field of machine learning. The MNIST contains $60,000$ training images and $10,000$ test images.

CIFAR-10 (Krizhevsky et al., 2009) contains $60,000$ color images with 10 classes. The training set has $50,000$ images and the test set has $10,000$ images.

ImageNet-100 is composed by randomly sampled 100 categories from ImageNet-1K (Deng et al., 2009). This dataset contains the following classes: n01498041, n01514859, n01582220, n01608432, n01616318, n01687978, n01776313, n01806567, n01833805, n01882714, n01910747, n01944390, n01985128, n02007558, n02071294, n02085620, n02114855, n02123045, n02128385, n02129165, n02129604, n02165456, n02190166, n02219486, n02226429, n02279972, n02317335, n02326432, n02342885, n02363005, n02391049, n02395406, n02403003, n02422699, n02442845, n02444819, n02480855, n02510455, n02640242, n02672831, n02687172, n02701002, n02730930, n02769748, n02782093, n02787622, n02793495, n02799071, n02802426, n02814860, n02840245, n02906734, n02948072, n02980441, n02999410, n03014705, n03028079, n03032252, n03125729, n03160309, n03179701, n03220513, n03249569, n03291819, n03384352, n03388043, n03450230, n03481172, n03594734, n03594945, n03627232, n03642806, n03649909, n03661043, n03676483, n03724870, n03733281, n03759954, n03761084, n03773504, n03804744, n03916031, n03938244, n04004767, n04026417, n04090263, n04133789, n04153751, n04296562, n04330267, n04371774, n04404412, n04465501, n04485082, n04507155, n04536866, n04579432, n04606251, n07714990, n07745940.

MNIST-C (Mu & Gilmer, 2019) is a corrupted version of MNIST data with different corruption types for benchmarking out-of-distribution robustness in computer vision.

CIFAR-10-C is algorithmically generated, following the previous leterature (Hendrycks & Dietterich, 2018), from different corruptions for CIFAR-10 data including gaussian noise, defocus blur, glass blur, impulse noise, shot noise, snow, and zoom blur.

ImageNet-100-C is algorithmically generated with Gaussian noise based on (Hendrycks & Dietterich, 2018) for the ImageNet-100 dataset (Deng et al., 2009).

FashionMNIST (Xiao et al., 2017) consists of $70,000$ fashion products images from 10 categories, with $7,000$ images per category. There are $60,000$ training images and $10,000$ test images. The 10 categories include T-Shirt, Trouser, Pullover, Dress, Coat, Sandals, Shirt, Sneaker, Bad, and Ankle boots.

SVHN (Netzer et al., 2011) is a real-world image dataset obtained from house numbers in Google Street View images. This dataset $73,257$ samples for training, and $26,032$ samples for testing with 10 classes.

Places365 (Zhou et al., 2017) contains scene photographs and diverse types of environments encountered in the world. The

Table 7. Additional results. Comparison with competitive OOD detection and OOD generalization methods on CIFAR-10. For experiments using $\mathbb{P}_{\text{wild}}$, we use $\pi_s = 0.5$, $\pi_c = 0.1$. For each semantic OOD dataset, we create corresponding wild mixture distribution $\mathbb{P}_{\text{wild}} := (1-\pi_s-\pi_c)\mathbb{P}_{\text{in}} + \pi_s\mathbb{P}_{\text{out}}^{\text{semantic}} + \pi_c\mathbb{P}_{\text{out}}^{\text{covariate}}$ for training.

Model	Texture $\mathbb{P}_{\text{out}}^{\text{semantic}}$, CIFAR-10-C $\mathbb{P}_{\text{out}}^{\text{covariate}}$				LSUN-R $\mathbb{P}_{\text{out}}^{\text{semantic}}$, CIFAR-10-C $\mathbb{P}_{\text{out}}^{\text{covariate}}$			
	OOD Acc.↑	**ID Acc.↑**	**FPR↓**	**AUROC↑**	**OOD Acc.↑**	**ID Acc.↑**	**FPR↓**	**AUROC↑**
OOD detection								
MSP	75.05	94.84	59.28	88.50	75.05	94.84	52.15	91.37
ODIN	75.05	94.84	49.12	84.97	75.05	94.84	26.62	94.57
Energy	75.05	94.84	52.79	85.22	75.05	94.84	27.58	94.24
Mahalanobis	75.05	94.84	15.00	97.33	75.05	94.84	42.62	93.23
ViM	75.05	94.84	29.35	93.70	75.05	94.84	36.80	93.37
KNN	75.05	94.84	39.50	92.73	75.05	94.84	29.75	94.60
OOD generalization								
ERM	75.05	94.84	52.79	85.22	75.05	94.84	27.58	94.24
Mixup	79.17	93.30	58.24	75.70	79.17	93.30	32.73	88.86
IRM	77.92	90.85	59.42	87.81	77.92	90.85	34.50	94.54
VREx	76.90	91.35	65.45	85.46	76.90	91.35	44.20	92.55
Learning w. $\mathbb{P}_{wild}$								
OE	44.71	92.84	29.36	93.93	46.89	94.07	0.7	99.78
Energy (w/ outlier)	49.34	94.68	16.42	96.46	32.91	93.01	0.27	99.94
Woods	83.14	94.49	39.10	90.45	78.75	95.01	0.60	99.87
Scone (ours)	85.56	93.97	37.15	90.91	80.31	94.97	0.87	99.79

scene semantic categories consist of three macro-classes: Indoor, Nature, and Urban.

LSUN-C (Yu et al., 2015) and **LSUN-R** (Yu et al., 2015) are large-scale image datasets that are annotated using deep learning with humans in the loop. LSUN-C is a cropped version of LSUN and LSUN-R is a resized version of the LSUN dataset, which has no overlap categories with the CIFAR dataset (Krizhevsky et al., 2009).

Textures (Cimpoi et al., 2014) refers to the Describable Textures Dataset, which contains a large dataset of visual attributes including patterns and textures. The subset we used has no overlap categories with the CIFAR dataset (Krizhevsky et al., 2009).

iNaturalist (Van Horn et al., 2018) is a challenging real-world dataset with iNaturalist species, captured in a wide variety of situations. It has 13 super-categories and 5,089 sub-categories. We use the subset from (Huang & Li, 2021) that contains 110 plant classes that no category overlaps with IMAGENET-1K (Deng et al., 2009).

Details of data split for OOD datasets. For datasets with standard train-test split (e.g., SVHN), we use the original test split for evaluation. For other OOD datasets (e.g., LSUN-C), we use 70% of the data for creating the wild mixture training data as well as the mixture validation dataset. We use the remaining examples for test-time evaluation. For splitting training/validation, we use 30% for validation and the remaining for training.

D. Results on Additional OOD Datasets

In this section, we provide the main results on more OOD datasets including Textures (Cimpoi et al., 2014) and LSUN_Resize (Yu et al., 2015) in Table 7. We observe that our proposed approach achieves overall strong performance in OOD generalization and OOD detection on these additional OOD datasets. Particularly, we compare our method with post hoc OOD detection methods such as `MSP` (Hendrycks & Gimpel, 2017), `ODIN` (Liang et al., 2018), `Energy` (Liu et al., 2020), `Mahalanobis` (Lee et al., 2018), `ViM` (Wang et al., 2022a), and `KNN` (Sun et al., 2022). These methods are all based on a model trained with cross-entropy loss, which suffers from limiting OOD generalization performance (75.05%). In contrast, our method achieves an improved OOD generalization performance (e.g., 85.56% when the wild data is a mixture of CIFAR-10, CIFAR-10-C, and Texture).

We also compare our method with common OOD generalization baseline methods including `IRM` (Arjovsky et al., 2019), `Mixup` (Zhang et al., 2018), and `VREx` (Krueger et al., 2021). Our method consistently achieves better results compared to

Algorithm	Art painting	Cartoon	Photo	Sketch	Average Acc. (%)
ERM (Vapnik, 1999)	88.1	77.9	97.8	79.1	85.7
IRM (Arjovsky et al., 2019)	85.0	77.6	96.7	78.5	84.4
GroupDRO (Sagawa et al., 2019)	86.4	79.9	98.0	72.1	84.1
I-Mixup (Wang et al., 2020; Xu et al., 2020)	86.5	76.6	97.7	76.5	84.3
VREx (Krueger et al., 2021)	86.0	79.1	96.9	77.7	84.9
MLDG (Li et al., 2018a)	89.1	78.8	97.0	74.4	84.8
CORAL (Sun & Saenko, 2016)	87.7	79.2	97.6	79.4	86.0
MMD (Li et al., 2018b)	84.5	79.7	97.5	78.1	85.0
DANN (Ganin et al., 2016)	85.9	79.9	97.6	75.2	84.6
CDANN (Li et al., 2018c)	84.0	78.5	97.0	71.8	82.8
MTL (Blanchard et al., 2021)	87.5	77.1	96.4	77.3	84.6
SagNet (Nam et al., 2021)	87.4	80.7	97.1	80.0	86.3
ARM (Zhang et al., 2021)	86.8	76.8	97.4	79.3	85.1
RSC (Huang et al., 2020)	85.4	79.7	97.6	78.2	85.2
Ours	88.5	83.8	96.2	77.3	86.4

Table 8. Comparison with OOD generalization algorithms on the PACS dataset from DomainBed benchmark. All methods are trained on ResNet-50. The model selection is based on a training domain validation set.

these OOD generalization baselines. Lastly, we compare our method with strong OOD detection methods using $\mathbb{P}_{\text{wild}}$ such as OE (Hendrycks et al., 2018), energy-regularized learning (Liu et al., 2020), and WOODS (Katz-Samuels et al., 2022). Our method demonstrates strong performance on OOD generalization accuracy, which shows the effectiveness of our method for making use of the covariate OOD data.

E. Results on PACS

In this section, we report results on the PACS dataset (Li et al., 2017) and present the comparisons with other baseline methods from the DomainBed. As shown in Table 8, we compare our method with a collection of common OOD generalization baselines, including `ERM` (Vapnik, 1999), `IRM` (Arjovsky et al., 2019), `GroupDRO` (Sagawa et al., 2019), `I-Mixup` (Zhang et al., 2018), `VREx` (Krueger et al., 2021), `MLDG` (Li et al., 2018a), `CORAL` (Sun & Saenko, 2016), `MMD` (Li et al., 2018b), `DANN` (Ganin et al., 2016), `CDANN` (Li et al., 2018c), `MTL` (Blanchard et al., 2021), `SagNet` (Nam et al., 2021), `ARM` (Zhang et al., 2021), `RSC` (Huang et al., 2020). Our approach SCONE (86.4%) outperforms all of these OOD generalization baselines on the DomainBed benchmark.

As shown in Table 9, we summarize not only the OOD generalization performance but also the OOD detection performance on the PACS dataset. The results indicate that SCONE displays strong performance on both OOD generalization and detection tasks.

Table 9. Results for both OOD generalization and detection tasks on the PACS dataset.

Method	OOD Accuracy	ID accuracy	FPR95	AUROC
Photo	96.23	99.68	2.57	99.38
Art painting	88.46	99.63	1.70	99.43
Cartoon	83.75	99.52	0.63	99.68
Sketch	77.25	99.01	17.80	96.34
Average	86.42	99.46	5.68	98.71

F. Results on Different Corruption Types

In this section, we provide additional ablation studies of the different covariate shifts. In Table 10, we evaluate our method under 19 different common corruptions such as *gaussian noise*, *defocus blur*, *glass blur*, *impulse noise*, *shot noise*, *snow*, *zoom blur*, *brightness*, etc. We follow the default design and parameter setting as in the original paper (Hendrycks & Dietterich, 2018) for generating the corruptions. For our method with margin, η is chosen based on the validation strategy in Appendix B. Our method is overall more robust under different covariate shifts than the WOODS baseline.

Table 10. Ablations on the different covariate shifts. We train on CIFAR-10 as ID, using CIFAR-10-C as $\mathbb{P}_{\text{ood}}^{\text{covariate}}$ and SVHN as $\mathbb{P}_{\text{ood}}^{\text{semantic}}$ (with $\pi_c = 0.5$ and $\pi_s = 0.1$).

Covariate shift type	Method	OOD Acc.↑	ID Acc.↑	FPR↓	AUROC↑
Gaussian noise	WOODS	52.76	94.86	2.11	99.52
Gaussian noise	SCONE	84.69	94.65	10.86	97.84
Defocus blur	WOODS	94.76	94.99	0.88	99.83
Defocus blur	SCONE	94.86	94.92	11.19	97.81
Frosted glass blur	WOODS	38.22	94.90	1.63	99.71
Frosted glass blur	SCONE	69.32	94.49	12.80	97.51
Impulse noise	WOODS	70.24	94.87	2.47	99.47
Impulse noise	SCONE	87.97	94.82	9.70	97.98
Shot noise	WOODS	70.09	94.93	3.73	99.26
Shot noise	SCONE	88.62	94.68	10.74	97.85
Snow	WOODS	88.10	95.00	2.42	99.54
Snow	SCONE	90.85	94.83	13.22	97.32
Zoom blur	WOODS	69.15	94.86	0.38	99.91
Zoom blur	SCONE	90.87	94.89	7.72	98.54
Brightness	WOODS	94.86	94.98	1.24	99.77
Brightness	SCONE	94.93	94.97	1.41	99.74
Elastic transform	WOODS	87.89	95.04	0.37	99.92
Elastic transform	SCONE	91.01	94.88	8.77	98.32
Contrast	WOODS	94.37	94.94	1.06	99.80
Contrast	SCONE	94.40	94.98	1.30	99.77
Fog	WOODS	94.69	95.01	1.06	99.80
Fog	SCONE	94.71	95.00	1.35	99.76
Frost	WOODS	87.25	94.97	2.35	99.55
Frost	SCONE	91.94	94.85	10.08	98.03
Gaussian blur	WOODS	94.78	94.98	0.87	99.83
Gaussian blur	SCONE	94.76	94.86	3.14	99.39
Jpeg	WOODS	84.35	94.96	1.73	99.68
Jpeg	SCONE	87.87	94.90	8.14	98.49
Motion blur	WOODS	82.54	94.79	0.47	99.88
Motion blur	SCONE	91.95	94.90	9.15	98.18
Pixelate	WOODS	91.56	94.91	1.82	99.66
Pixelate	SCONE	92.08	94.96	1.97	99.64
Saturate	WOODS	92.45	95.03	1.26	99.77
Saturate	SCONE	93.38	94.92	10.27	97.88
Spatter	WOODS	92.38	94.98	1.94	99.64
Spatter	SCONE	92.78	94.98	1.94	99.64
Speckle noise	WOODS	72.31	94.94	3.51	99.30
Speckle noise	SCONE	88.51	94.83	11.05	97.82