Machine Learning
https://doi.org/10.1007/s10994-022-06237-1

Open-set learning under covariate shift

Jie-Jing Shao[1] · Xiao-Wen Yang[1] · Lan-Zhe Guo[1]

Received: 26 May 2022 / Revised: 1 September 2022 / Accepted: 12 September 2022

Abstract

Open-set learning deals with the testing distribution where there exist samples from the classes that are unseen during training. They aim to classify the seen classes and recognize the unseen classes. Previous studies typically assume that the marginal distribution of the seen classes is fixed across the training and testing distributions. In many real-world applications, however, there may exist covariate shift between them, i.e., the marginal distribution of seen classes may shift. We call this kind of problem as *open-set learning under covariate shift*, aim to robustly classify the seen classes under covariate shift and be aware of the unseen classes.We present a new open-set learning framework with covariate generalization based on supervised contrastive learning, called SC–OSG, inspired by the latent connection between contrastive learning and representation invariance. Specifically, we theoretically justify supervised contrastive learning that could promote the conditional invariance of representations, a critical condition for covariate generalization. SC–OSG generates multi-source samples to promote the representation invariance and improve the covariate generalization. Based on this, we propose a detection score that is specific to the proposed training scheme. We evaluate the effectiveness of our method on several real-world datasets, on all of which we achieve competitive results with state-of-the-art methods.

Keywords Open-set learning · Out-of-distribution generalization · Representation learning · Image classification

Communicated by Yu-Feng Li, Prateek Jain.

Jie-Jing Shao and Xiao-Wen Yang have contributed equally to this work.

✉ Lan-Zhe Guo
guolz@lamda.nju.edu.cn

Jie-Jing Shao
shaojj@lamda.nju.edu.cn

Xiao-Wen Yang
yangxw@lamda.nju.edu.cn

[1] National Key Laboratory for Novel Software Technology, Nanjing University, Nanjing 210023, China

Published online: 27 October 2022

1 Introduction

Supervised learning has achieved competitive or even better performance than human beings in a variety of tasks, like image understanding and language processing (LeCun et al., 2015). However, they typically rely on the basic assumption, i.e., testing distribution is the same as training distribution. When a model is deployed in the open world, such a condition is difficult to satisfy, as the testing samples may arise from unseen classes (Geng et al., 2021). For example, in document classification (Fei and Liu, 2016), irrelevant documents may appear in the testing data and lead to a misprediction. Similar cases commonly appear in other applications, such as self-driving (Wong et al., 2019) and activity recognition (Yang et al., 2019). In such applications, misprediction of unseen class may lead to risk, like traffic accidents in self-driving scenarios. In order to deal with this kind of problem, open-set learning has been proposed and attracted considerable attention, which has consequently resulted in a large number of open-set learning methods (Da et al., 2014; Bendale and Boult, 2016; Yoshihashi et al., 2019; Tack et al., 2020; Vaze et al., 2022).

Open-set learning is proposed to not only classify seen classes but also recognize the unseen classes. They attempt to detect samples that do not belong to the training distribution, by exploring different strategies, such as open-space risk (Da et al., 2014; Zhou et al., 2021a), extreme value theory (Bendale and Boult, 2016; Yoshihashi et al., 2019), representation learning methods (Tack et al., 2020; Winkens et al., 2020), and other interesting techniques (Geng et al., 2021). Recently, some works (Tack et al., 2020; Vaze et al., 2022) indicated that whether representations have sufficient discrimination on closed-set is critical for open-set performance. All of the above positive results, however, are based on a basic assumption that the marginal distribution of known class is fixed in the training and testing distribution. Such an assumption is difficult to hold in many real-world applications where testing distribution may shift from the training distribution, such as self-driving (Yu et al., 2020), influenza detection (Rejmanek et al., 2015), and speech recognition (Liao, 2013). These methods which designed for detecting samples that do not belong to the seen distribution, become unreliable when the marginal distributions of seen classes shifts.

Out-of-distribution generalization (domain generalization) aims at generalizing the model to covariate shift where the marginal distribution shifts from the training to the testing phase. Most of them attempted to mine the stable relationship across multiple sources, like marginal distribution alignment (Ganin and Lempitsky, 2015; Kim et al., 2021) or conditional distribution alignment (Arjovsky et al., 2019; Ahuja et al., 2020). However, they typically assume there are multiple training sources. It can hardly be satisfied, where modern datasets are frequently assembled without explicit source labels. Recently, some works have considered the general setting where prior division is not available. They divided the training samples into some subsets to mine the latent heterogeneity (Liu et al., 2021; Creager et al., 2021; Zhang et al., 2021). Nevertheless, we find they still perform poorly for open-set learning.

To summarize, we consider the *open-set learning under covariate shift* problem. More formally, given features X and labels Y, we observe training distribution $P^{tr}(X^{tr}, Y^{tr})$, and deploy the model in the testing distribution $P^{te}(X^{te}, Y^{te})$. 1) There may exist sample $(x, y) \sim P^{te}$, where unseen class $y \in Y^{te}$ has not appeared in the training $y_i \notin Y^{tr}$. 2) The marginal distribution of seen class may shift $P^{te}(x|y) \neq P^{tr}(x|y), y \in Y^{tr} \subset Y^{te}$. In this paper, we call them *semantic shift* and *covariate shift* respectively. Figure 1 has illustrated this problem setting.

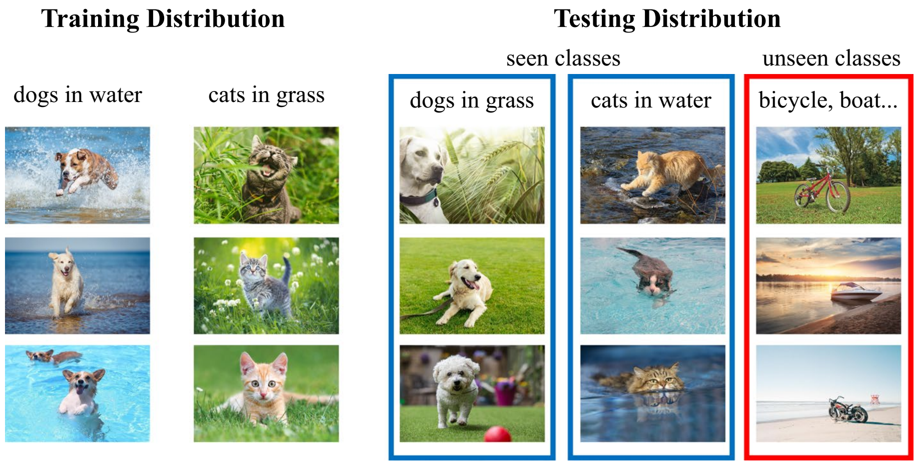

Fig. 1 Open-set learning under covariate shift. There are both covariate shift (blue) and semantic shift (red) in the testning distribution

In this work, we present a new open-set learning framework with covariate generalization based on supervised contrastive learning, called SC–OSG, inspired by the latent connection between contrastive learning and representation invariance. We first indicate that supervised contrastive learning, which has achieved empirical success in i.i.d. representation learning scenarios, could promote the conditional invariance of representations. Moreover, we propose a theoretically grounded framework to learn invariant representation by synthetic multi-source samples. We next take a feature attribution method to decompose and re-generate samples as an implementation of our framework. Based on this, we propose a detection score that is specific to the proposed training scheme. We demonstrate the effectiveness of our method on several real-world datasets, on all of which we achieve competitive results with state-of-the-art methods.

2 Related work

2.1 Open-set learning

Open-Set Learning has been studied for a long history to handle the semantic shift. Recently, deep learning-based methods have attracted much attention due to their powerful representations. They could be mainly divided into three schemes. The first assumes there are unlabeled data from testing distribution available, which contains both known classes and unknown classes (Da et al., 2014; Liu et al., 2018; Yu and Aizawa, 2019; Zhou et al., 2021). They detect the outliers from the unlabeled data and utilize them to build an open-set classifier. The second assumes the samples from unknown classes are unavailable and anticipates the novel classes via generative models (Neal et al., 2018; Fang et al., 2021; Chen et al., 2021). The third considers representation learning and improving separation between known classes (Hendrycks et al., 2019; Winkens et al., 2020; Tack et al., 2020; Vaze et al., 2022).

Recently, there are some work introduce open-set unlabeled data (Huang et al. 2021, 2022) and utilize representation transfer to boost the performance. Nevertheless, all of these promising results assume the distribution of known classes is fixed across training and testing distributions or the open-set shift is available at training.

2.2 Out-of-distribution generalization

Out-of-distribution generalization has recently attracted much attention in handling the covariate shift. They mainly mine the stable relationship across multiple sources, like marginal distribution alignment (Ganin and Lempitsky, 2015; Kim et al., 2021) or conditional distribution alignment (Arjovsky et al., 2019; Ahuja et al., 2020). Ganin and Lempitsky (2015) build a domain discriminator to adversarially align the cross-source marginal distribution. Kim et al. (2021) introduces contrastive loss and designs a class-specific domain perturbation layer to extract the domain-invariant representations. Arjovsky et al. (2019) and Ahuja et al. (2020) learn the invariant conditional relationship to obtain a stable learner. However, they typically assume there is a prior division of multiple sources which can hardly be satisfied, modern datasets are frequently assembled by merging data from multiple sources without explicit source labels. Recently, some work attempted to mine the latent heterogeneity without prior source division (Liu et al., 2021; Creager et al., 2021; Zhang et al., 2021). In contrast, we consider enerating auxiliary distributions to enrich the diversity rather than dividing the raw samples into sub-sources.

There are some previous works focused on both semantic shift and covariate shift, like open-set domain adaptation (Busto and Gall, 2017; Baktashmotlagh et al., 2019; Luo et al., 2020) and Open-Set Domain Generalization (Shu et al., 2021). Open-Set Domain Adaptation assumes the unlabeled data from the target testing distribution is available. The existing Open-Set Domain Generalization work (Shu et al., 2021) assumes there is a prior domain division. This condition is different from ours which exploits the training data without domain labels.

Their solutions have not well addressed the problem we studied.

2.3 Contrastive learning

Contrastive learning has shown remarkable success in visual representation learning (Bachman et al., 2019; Hjelm et al., 2019; Chen et al., 2020; Tian et al., 2020; Hendrycks et al., 2019; Tack et al., 2020), which encourages closer representations for augmentations (views) of the same natural data than for randomly sampled pairs of data. Inspire by these findings, we introduce the deviation of contrastive learning when supervised signals are available, supervised contrastive learning (SCL) (Khosla et al., 2020), to learn the invariant representations. In this paper, we theoretically find the connection between SCL and covariate generalization. Furthermore, we build an efficient framework SC–OSG for the open-set learning under covariate shift problem.

3 SCL on representation invariance

3.1 Problem formulation

Following structural causal model (Arjovsky et al., 2019; Ahuja et al., 2020; Creager et al., 2021), let us consider the data generation process epicted in Fig. 2. In this graphical model, the label Y and environment E are generated first from their prior distribution $P(Y)$ and $P(E)$.

The input features X are the observation of semantic and environmental variables (Y, E). This generation process can be exploited in many classification applications. Taking the images as examples, Y is the semantic factor (e.g., the shape of an object) and E is the environmental factor (e.g., background, object position). Due to the correlation $P^{tr}(Y, E)$ from training data (e.g., cars often appear in a road background and cats on the grass), the model learned on training distribution may rely on the E for prediction via this correlation. Note such a correlation is not stable in different distributions $P^e(X, Y)$ and is just a training-specific property. In unknown testing distributions, such a correlation is not reliable. As the graphical model indicated, in an arbitrary testing distribution, $P(Y, E) = P(Y)P(E)$, i.e., the semantic factors and environmental factors are independent.

Take the example of Fig. 1, we could find cats (Y) always appeared in the grass (E) in the training data. Such a training-specific correlation $P(Y, E)$ will lead the model to cut corners, i.e., make a prediction “cat" when it observes the grass background. It brings challenges for both seen class classification and unseen class detection, e.g., the model tends to predict a cat label when it encounters a dog or a bicycle on the grass. Here we give the formal assumption on different distributions under the covariate shift condition.

Assumption 1 The feature X is generated by an unknown process $G(E, Y)$ of two independent factors, semantics Y and environments E. We assume there are optimal representations $Z = g^*(X)$ such that the following properties hold:

1) Invariance property: we have the conditional correlation $P^e(Y|Z) = P^{e'}(Y|Z)$ holds, $\forall e, e' \in E$.
2) Sufficiency property: $Y = h(Z)$ across distributions $P^e(X, Y)$.

the training and testing distribution arise from the joint distribution $P^{tr}(X, Y), P^{te}(X, Y) \sim X \times Y$. There are semantic shift $P^{tr}(y) = 0 < P^{te}(y), \exists y \in Y$ and marginal shift $P^{tr}(x|y) \neq P^{te}(x|y), \forall y \in Y^{tr} \subset Y^{te}$.

Assumption 1 indicates invariance and sufficiency for identifying semantic differences using representation Z which has stable relationships with Y across different distributions $P^e(X, Y), \forall e \in E$.

Here, we first introduce the definition and properties of conditional invariance and then show that SCL, an empirically successful framework in the i.i.d. scenario, could decouple the correlation between semantics Y and environments E, promoting representations to conditional invariance.

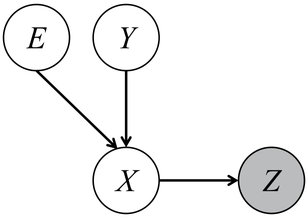

Fig. 2 Graphical model. $P^e(X, Y)$ is the data distribution for environment e. Our goal is to learn representations Z from X which can be generalized across distributions

3.2 Representation invariance

There is a branch of work on out-of-distribution generalization that attempts to learn the invariant representation, decoupling the spurious correlation between environments and semantics to improve the robustness of unseen testing distribution. In this paper, we focus on the conditional invariance:

Definition 1 Let $Z = g(X)$ be the representation distribution via learning module g. If we have $P^e(Z|Y) = P^{e'}(Z|Y), \forall e, e' \in E$, the representation Z has achieved conditional invariance for environments E.

Remark 1 Under Assumption 1, if representations Z satisfy the conditional invariance on environments E, and classifier h achieves ideal risk on environment e, i.e., $\mathbb{E}^e[R(h(Z), Y)] \rightarrow 0, \exists e \in E$. We have $\mathbb{E}^{e'}[R(h(Z), Y)] \rightarrow 0, \forall e' \in E$.

Intuitively, in different distributions, $P^e(X, Y), e \in E$, the conditional invariant representation Z could embed samples $P^e(X|Y)$ from any class Y to the same representation distribution, and excluding the influence of environmental-related spurious correlation. It indicates that we can guarantee the generalization to different environments as long as the conditional invariance of the representation holds.

As we discussed above, most of the existing works need to collect multiple training sources and learn the invariant representation, which limits their applications on open-set learning. When we do not have multiple training sources, how to improve the generalization has been a quite challenging problem.

We find supervised contrastive learning (SCL) is a potential solution to address it. SCL is a variant version of contrast learning when supervised signals are available (Khosla et al., 2020). It consists of two parts, data augmentation $\mathcal{A}$ (e.g. rotation, flipping, grey-scaling for images), and contrastive loss optimization. Formally, the supervised contrastive learning SC(g;P, $\mathcal{A}(P)$) could be formulated as:

$$\min_{g} \sum_{(x_i,y_i)\sim P(X,Y)} \frac{-1}{|P(y_i)|} \sum_{x_j \in P(y_i)} \log \frac{\exp(g(x_i)\cdot g(\mathcal{A}(x_j))/\tau)}{\sum_{x_k \in \mathcal{X}} \exp(g(x_i)\cdot g(\mathcal{A}(x_k))/\tau)} \tag{1}$$

where $P(y_i)$ is the set of samples from the class y_i, and $\mathcal{A}$ represents the specific data augmentation. Following (Zimmermann et al., 2021; HaoChen et al., 2021; Wen and Li, 2021), we assume the augmentation operation does not change the semantic factors from raw input distribution X, and regard $\mathcal{A}(X)$ as an auxiliary distribution. We then theoretically justify that the learning of SCL (Eq. 1) aligns the conditional distribution of representation Z.

Theorem 1 *When data augmentation $\mathcal{A}$ does not change the semantic causes of inputs, we regrad the raw distribution $P^{\text{raw}}(X, Y)$ and the augmented distribution $P^{\text{aug}}(\mathcal{A}(X), Y)$ as two distributions. Then, the supervised contrastive learning (as Eq. 1) maximizes the mutual information between $P^{\text{raw}}(Z|Y)$ and $P^{\text{aug}}(Z|Y)$ and promotes the conditional invariance.*

Proof provided in the appendix. □

Taking the Theorem 1 and the benefit from conditional invariance, we could justify that SCL is generalizable to testing distribution under marginal shift $\mathcal{A}$. Specifically, the learned model is robust to specific style transformation (e.g., change of object position, lighting). Such a finding also explains why SCL achieves an empirical improvement over cross-entropy-based methods in the i.i.d. scenario (Khosla et al., 2020; Chuang et al., 2020; Tian et al., 2020a).

We could find the generalization of SCL relies on the construction of the auxiliary domain, i.e. the design of data augmentation. The ideal data augmentation should enable as many environmental changes as possible without changing the semantic information. Inspired by the data generation process, we have the following proposition:

Proposition 1 *Through augmenting synthetic samples from different environments, supervised contrastive learning could promote the representation invariance under different environments. Formally, it alleviates the training-specific correlation $P^{tr}(Y, E)$ and simulate a distribution where $P(Y, E) = P(Y)P(E)$.*

Remark 2 When representations $Z = g(X)$ satisfy the conditional invariance between the raw domain and the ideal synthetic domains $P^{tr}(Z|X) = P^{e}(Z|X), \forall e \in E$, and there exists h with low risk on raw distribution $R^{raw}(h(Z), Y) \rightarrow 0$, for any $e \in E$ they could achieve the consistent low risk $R^{e}(h(Z), Y) \rightarrow 0$.

It indicates that, if we can build the functions that disentangle the semantic and environmental factors, and re-generate the distribution $P(Y, E) = P(Y)P(E)$, we could promote the conditional invariance between the raw domain and the augmented domain via SCL. Then the learned model could be generalizable to any distribution P^{e}. Take the case of Fig. 1 again, we attempt to sample "cats with water" and "dogs with grass". Then, we could embed the same semantic samples from different environments to the same representation distributions, and improve the generalization through the conditional alignment framework. The overall framework becomes:

$$\min_{g,h} R(h \circ g; P^{\mathrm{raw}}(X, Y)) + \sum_{e \in E} \mathrm{SC}(g; P^{\mathrm{raw}}, P^{e}), \text{ s.t. } P^{e} = \mathcal{A}^{e}(P^{\mathrm{raw}}) \tag{2}$$

where the first term represents a classification risk measure and the second term implements the conditional invariance between P^{raw} and P^{e} via the SCL, i.e., the specific data augmentation $\mathcal{A}^{e}$ transforms the raw distribution to P^{e}.

As discussed above, we would like to build such an auxiliary environment for our invariant representation learning framework. In the next section, we show how one can incorporate this idea into real-world generalization problems by constructing an auxiliary environment with feature attribution.

4 The proposed framework

4.1 Training phase

In the training phase, we consider constructing auxiliary environments and learning the environmental-invariant representations via data augmentation and SCL respectively. As we discussed above, an ideal implementation is to decompose the semantic and environmental factors from training data $P^{tr}(X, Y)$. Unfortunately, such a disentanglement is difficult to achieve without external assumptions about the data structure, especially for high-dimensional inputs like images.

In our implementation, we take the feature attribution of model inference in replacement of the latent causal factor. Feature attribution (Ribeiro et al., 2016; Lundberg & Lee, 2017; Selvaraju et al., 2020) is a class of methods that focuses on finding which feature results in the specific model inference. Generally, given a sample (x, y), feature attribution attempts to estimate the influence of each $[x^1, x^2, \dots x^d]$ on model inference $f(x) = y$, where d is the dimension of input x. More formally, they would obtain scores $\{s(x_i^k)\}_{k=1}^d \in [0, 1]$ for each pixel x_i^k of input x_i, which measure their influence on the model inference $f(x_i) = y_i$. Specifically, we take the Grad-CAM (Selvaraju et al., 2020), a well-known feature attribution technique, to catch the cause justification from the model. Grad-CAM uses the gradients of any target concept (say "dog" in the classification model) flowing to produce a localization map on x highlighting the important regions in the image. As shown in Fig. 3, we could obtain the pixel-level feature attribution $s(x_i) \in \mathbb{R}^d$ via Grad-CAM. Although there are different feature attribution methods to calculate scores. In this paper, we mainly focus on the Grad-CAM, because it is model-architecture-independent and easy to implement.

Then we remix these semantics factors with environmental factors from different samples. Inspired by mixup (Zhang et al., 2018), we generate the diverse environmental part via a convex combination.

Given two samples $(x_i, y_i), (x_j, y_j)$, we could generate an auxiliary sample $\hat{x}_{i,j} = \text{EnvMix}(x_i, x_j)$ which is composed with the semantic part from x_i and environmental part from x_j, as the following:

$$\{s(x_i^k)\}_{k=1}^d = \text{Grad-CAM}(x_i, y_i) \tag{3a}$$

$$c_i = [c_i^k], c_i^k = x_i^k \cdot \mathbb{I}(s(x_i^k) > \gamma)] \tag{3b}$$

$$e_i = [e_i^k], e_i^k = x_i^k \cdot \mathbb{I}(s(x_i^k) \leq \gamma) + \mathcal{N}(\mu_i, \sigma_i) \cdot \mathbb{I}(s(x_i^k) > \gamma)] \tag{3c}$$

$$\hat{e}_{i,j} = \lambda e_i + (1 - \lambda) e_j \tag{3d}$$

$$\hat{x}_{i,j} = c_i + \hat{e}_{i,j} \tag{3e}$$

We first obtain the attribution score via Grad-CAM and then split the image x_i into the class-dependent part c_i and the environmental part e_i on pixel-level, as Eqs. 3a, 3b and 3c.

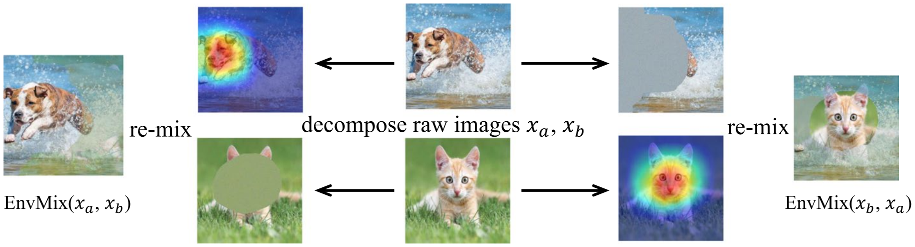

Fig. 3 EnvMix. We decompose the factors via Grad-CAM and generate synthetic samples

In Eq. 3c, we use Gaussian noise to fill in the gaps caused by stripping the semantic part, where μ_i and σ_i represent the mean and standard deviation of raw x_i. After the same process, we can get the environmental part e_j of the sample x_j. In Eq. 3d, we generate a new background via a convex combination of e_i and e_j, where $\lambda \sim Beta(1, 1)$ is a mixing parameter. Finally, we combine the c_i and the generated $\hat{e}_{i,j}$, which samples the semantics c_i from the synthetic environment. The Fig. 3 has demonstrated a case from our generation process.

Through such a data generalization process, we sample the semantic samples Y from each $e \in E$, which are extracted from the training data via the feature attribution of model inference.

Although such a generation scheme could obtain sufficient samples from various environments, it costs much on data and computational complexity. Let the N be the size of the raw data, we would have N^2 samples from N environments. This is not affordable for the contrastive framework as the Eq. 2, in practice. Thus, we generate the synthetic samples in a batch mode, i.e., for a data batch $Q : Q[i] = x_i$ with size n, we get its random permutation Q', and then generate n samples via $\text{EnvMix}(Q[i], Q'[i])$. Then we take the generated samples as a comprehensive auxiliary domain. The pairwise mode of Eq. 2 could be reduced to the raw 2-view mode, like Eq. 1. Overall, we could execute the model training via a 3-forward process, which is more efficient in practice.

4.2 Testing phase

In the testing phase, we would like to identify if a sample arises from unseen classes and classify the seen classes. Following Bendale and Boult (2016), we first obtain the probability of the seen classes and reject the unseen classes based on a threshold ϵ. Especially, we estimate the class probability of x_i:

$$\hat{P}(y = j|x_i) = \frac{1}{2}\left(P(y = j; x_i, h \circ g) + \frac{\exp(g(x_i) \cdot \mathcal{P}_j / \tau)}{\sum_k \exp(g(x_i) \cdot \mathcal{P}_k / \tau)} \right) \tag{4}$$

where $\mathcal{P}_j = 1/|P(j)| \sum_{x \in P(j)} g(x)$ is the expected representation of class j from training samples. In words, the Eq. 4 consists of two parts, a classification score, and a conditional representation distance, being consistent with our training scheme. The overall SC–OSG is summarized in Algorithm 1.

Algorithm 1 The overall framework (SC-OSG)

```
Given training data $D \sim P^{tr}(X, Y)$
for $t = 1$ to $T$ do
    Sample data $Q = [x_i]_{i=1}^{B}$ from $D$ with size $B$
    $Q' = \text{Shuffle}(Q)$
    Generate synthetic samples $\hat{Q} = [\hat{x}_i]_{i=1}^{B}, \hat{x}_i = \text{EnvMix}(Q[i], Q'[i])$
    Update model $h \circ g$ via supervised cross-entropy loss.
    Update the represtion module $g$ via $\text{SC}(g;\ Q, \hat{Q})$, as Equtation 1
end for
Given testing input $x$.
Get the probability score $\hat{P}(y|x)$ for known classes $y \in Y^{tr}$, as Equation 4.
Let $y^* = \arg\max_j \hat{P}(y = j|x)$
Identify as unseen class if $\hat{P}(y^*|x) < \epsilon$, otherwise $y^*$.
```

5 Empirical study

5.1 Experimental setup

To evaluate our method, we conduct experiments on three datasets that are commonly used in the literature for out-of-distribution generalization.

ImageCLEF-DA (Caputo et al., 2014) is a benchmark for ImageCLEF 2014 challenge which is widely used in domain adaptation and domain generalization. It contains four environments (collecting sources): *Caltech256* (**C**), *ImageNet ILSVRC2012* (**I**), *Pascal VOC 2012* (**P**) and *Bing* (**B**). For each environment, there are totally 600 images and 12 classes. We split the first 8 categories of environment **C** and **B** are divided as training set. The testing data consists of environments **I** and **P**, 4 out of its 12 classes did not appear in the training set.

NICO (He et al., 2021) dataset contains 19 classes, 188 environments, and nearly 250,000 images. Compared with the traditional benchmark as ImageCLEF-DA, there is no consistent source division, i.e., different classes of samples come from different sources, which are more frequent with real-world tasks and bring greater challenges to performance generalization. In this task, we split 10 animal classes (e.g., dog, cat, bird) to construct an animal recognition task, the rest 9 classes (bicycle, boat, bus, and so on) are regarded as the unknown classes. Each animal contains 10 environments like *in forest* or *on grass*. Three environments for each category are randomly split as training data, while others as testing data.

WILDS-FMoW (Christie et al., 2018) is a recently proposed out-of-distribution generalization benchmark for global-scale monitoring tasks, which consists of over 1 million images from 200 countries. These images are collected from 2002 to 2017, with distribution shifts over time. There are 80 environments and 63 categories in this dataset. As officially recommended, we divide the samples collected before 2013 into training environments and those after 2016 as test environments. We further split the half categories (31) as known classes, and the others will only appear in the testing data.

Competing methods We firstly consider the previous open-set learning methods: including three baselines Softmax (Hendrycks and Gimpel, 2017), OpenMax (Bendale and Boult, 2016), ODIN (Liang, et al., 2018) and 4 SOTA methods: ARPL (Chen et al., 2021), ARPL+CS (Chen et al., 2021), PROSER (Zhou et al., 2021a) and CTooD (Winkens et al., 2020). We also consider the previous out-of-distribution generalization methods, including 2 traditional methods: DANN (Ganin and Lempitsky, 2015) and MixupDG (Wang et al., 2020), which rely on the prior environmental divisions, and 2 SOTA division-free generalization methods: SelfReg (Kim et al., 2021; Zhang et al., 2021). We add a post-processing module, as a baseline solution to help them detect unknown classes (Hendrycks and Gimpel, 2017). Moreover, we compare with the DAML (Shu et al., 2021), the open-set domain generalization work, which is mostly related to us, although it needs a prior division for multiple training sources. To show the effectiveness of our proposal, we also conduct the experiments about the Mixup Zhang et al. (2018) and CutMix (Yun et al., 2019), two popular data augmentation methods, which are related to our EnvMix. To make an ablation comparison, we drop the two sub-modules, EnvMix and score function respectively. Specifically, SC–OSG† consists of basic SCL and the proposed score function, SC–OSG‡ consists of SCL with EnvMix and a soft-max score function. We also compare with the SCL, the most basic implementation for our framework.

Evaluation metrics We first consider the robustness under covariate shift. We use accuracy (ACC) on seen classes to evaluate if the model could recognize them when the marginal distribution shifts. We then consider the semantic shift. Following (Liang, et al., 2018; Chen et al., 2021), we evaluate the detection of unseen classes via a threshold-free metric: the area under the receiver operating characteristic curve (AUROC). Furthermore, we consider a comprehensive metric: open set classification rate (Dhamija et al., 2018) (OSCR) to measure performance on both seen and unseen classes. A high OSCR score indicates that classification and novelty detection receive excellent performance simultaneously. We provide its calculation in the appendix, a more in-depth discussion could be found in (Dhamija et al., 2018).

Implementation details For all of these methods, we use ResNet-18 (He et al., 2016) as the base model, which has been pre-trained on the ImageNet (Deng et al., 2009). Moreover, we employ SGD as the optimizer for all methods and we take the results when their losses converge. For our methods, we set γ of Env-Mix (Eqs. 3b 3c) to be the median of Grad-CAM scores by default. In words, it chose half of the input as the environment part while the other half as the semantic part. We then set a coordination coefficient to balance the classification loss and supervised contrastive loss, which makes the proportion of supervised contrastive loss increase linearly with the first few epochs and then fixed. In our experiment, we set the final coefficient as 0.1 both in the training and testing phases. Besides, all of the methods are implemented based on Pytorch and we train them on an NVIDIA RTX 3090.

5.2 Results analysis

In Table 1, we report the results on ImageCLEF-DA and NICO. In Table 2, we report the results on WILDS-FMoW. All of these results are obtained via three random repeats. From the results, we could find that out-of-distribution generalization methods provide a strong baseline for open-set learning under covariate shift, which even outperforms the SOTA open-set methods. Note that DANN, MixupDG, and DAML need prior division for multiple sources, so they cannot be used for datasets like NICO that do not have a consistent source division. An interesting finding is SelfReg which utilizes contrastive learning without the need for prior division, showing a strong performance. This is consistent with our finding that SCL is beneficial for distribution generalization. Nevertheless, our framework SC–OSG with EnvMix has shown a clear performance improvement, verifying its effectiveness.

5.3 Invariance comparison for SCL

In this work, we theoretically justify the effectiveness of SCL on covariate generalization. Here, we empirically compare the previous invariant representation learning methods, like DANN (Ganin and Lempitsky, 2015) and cross-entropy-based learning (CE). Specifically, we regard the raw data and the augmented data as two domains and learn the invariant representation of them.

Table 1 ACC, AUROC, and OSCR (%, mean ± std) on ImageCLEF-DA and NICO

	ImageCLEF-DA			NICO		
Method	ACC	AUROC	OSCR	ACC	AUROC	OSCR
Softmax	82.1±0.5	70.5±0.7	62.8±0.6	77.9±0.3	74.0±1.1	63.7±1.2
Openmax	82.1±0.5	68.8±0.6	50.8±1.4	77.9±0.3	74.3±1.7	61.4±0.8
ODIN	82.1±0.5	71.9±0.8	62.9±0.9	77.9±0.3	82.1±2.2	67.6±1.7
ARPL	74.5±0.8	72.7±2.7	59.5±2.3	74.9±1.0	80.7±1.7	65.5±0.5
ARPL+CS	63.3±1.2	62.9±1.5	45.3±1.9	71.3±1.4	74.7±2.0	59.1±1.7
PROSER	77.5±0.9	67.6±0.3	57.6±0.3	65.2±0.8	75.3±2.7	49.9±1.1
CTooD	79.5±0.4	69.8±0.9	60.9±0.3	73.3±0.4	72.8±0.9	59.4±0.7
DANN	82.5±0.6	69.6±1.1	62.1±1.0	-	-	-
MixupDG	81.6±0.4	70.9±0.7	62.3±0.6	-	-	-
DAML	81.3±0.8	69.0±0.5	61.0±0.5	-	-	-
SelfReg	82.2±0.3	72.4±0.1	64.1±0.5	**80.1±0.4**	80.0±1.2	69.3± 0.9
StableNet	80.6±1.1	69.7±1.5	61.8±1.3	74.2±0.4	70.8±1.4	59.2±1.3
Mixup	79.0±0.6	68.1±2.7	59.1±2.5	71.3±0.8	63.7±1.9	51.2±1.9
CutMix	79.5±0.9	64.2±2.2	55.5±2.2	67.6±0.8	68.3±2.6	52.4±2.2
SCL	82.2±0.6	72.2±0.8	64.1±0.8	79.8±0.5	76.7±0.8	66.7±0.6
SC–OSG†	84.1±0.6	77.7±0.9	68.2±0.7	79.3±0.5	87.4±0.3	**74.3±0.5**
SC–OSG‡	83.2±1.2	72.9±1.1	65.1±1.5	79.2±0.7	77.5±2.5	66.7±1.9
SC–OSG	**84.2±0.6**	**78.4±0.6**	**68.8±0.6**	78.9±0.6	**88.0±1.2**	74.2±0.6

Table 2 ACC, AUROC and OSCR (%, mean ± std) on WILDS-FMoW

Method	ACC	AUROC	OSCR	Method	ACC	AUROC	OSCR
Softmax	53.8±0.3	63.3±0.6	41.2±0.4	ARPL+CS	53.6±0.2	63.5±0.5	40.6±0.3
Openmax	53.8±0.3	63.2±0.5	40.4±0.4	PROSER	39.4±1.1	57.6±0.1	28.9±1.0
ODIN	53.8±0.3	64.2±0.7	40.0±0.6	CTooD	51.4±0.8	62.5±0.7	39.0±1.0
ARPL	51.1±0.8	63.2±0.9	38.6±1.2				
DANN	50.2±0.5	63.2±0.5	38.8±0.3	SelfReg	53.2±0.3	64.1±0.3	41.7±0.3
MixupDG	50.6±0.7	62.7±0.2	39.0±0.6	StableNet	**54.1±0.2**	64.2±0.2	41.9±0.3
Mixup	50.9±0.8	62.5±0.3	39.0±0.3	CutMix	50.9±1.2	63.1±0.6	38.5±0.9
SCL	52.0±0.5	62.6±0.9	39.9±0.7	SC–OSG‡	53.1±0.3	63.1±0.5	40.7±0.5
SC–OSG†	53.2±0.5	65.1±0.4	40.3±0.5	SC–OSG	53.7±0.2	**65.8±0.7**	**42.3±0.5**

From the results shown in Table 3, we could first find that invariant representation learning could outperform the CE baseline, which ignores the diversity among environments. Moreover, we could find our proposed SC–OSG could achieve a significant improvement, which verifies SCL as an effective framework for representation invariance.

Table 3 Performance comparison for different representation learning with EnvMix

Method	ACC	AUROC	OSCR
EnvMix+CE	81.3±0.8	69.5±2.3	61.9±2.1
EnvMix+DANN	82.3±0.1	71.4±1.2	63.5±1.1
SC–OSG	**83.2±1.2**	**72.9±1.1**	**65.1±1.5**

Table 4 ACC, AUROC, and OSCR (%, mean ± std) in traditional open-set setting

	ImageCLEF-DA			NICO		
Method	ACC	AUROC	OSCR	ACC	AUROC	OSCR
Softmax	86.3±0.6	77.2±0.4	71.8±0.6	93.3±0.5	87.2±0.6	83.9±0.5
Openmax	86.3±0.6	74.8±0.2	63.4±0.9	93.3±0.5	87.8±1.0	81.3±0.5
ODIN	86.3±0.6	77.2±1.0	71.2±1.1	93.3±0.5	92.6±1.2	87.4±0.7
ARPL	82.8±1.4	78.5±1.2	71.5±1.5	91.9±1.1	91.7±1.3	85.9±0.9
ARPL+CS	77.5±2.0	73.3±2.0	62.9±2.0	89.6±0.4	87.5±1.3	80.8±1.3
PROSER	81.5±1.3	67.7±0.9	66.6±0.4	83.3±0.8	83.9±1.8	71.4±1.2
CTooD	84.3±0.8	77.0±0.4	70.5±0.6	91.2±0.3	86.2±0.5	81.6±0.3
SCL	85.8±0.6	78.8±0.8	73.2±0.7	93.4±0.4	88.5±0.4	85.0±0.4
SC–OSG†	86.0±0.5	83.1±0.3	76.8±0.3	**93.6±0.1**	95.6±0.2	**92.0±1.3**
SC–OSG‡	**86.3±0.6**	78.6±1.0	73.2±1.1	92.8±0.5	89.0±2.3	85.2±2.3
SC–OSG	86.2±0.3	**83.4±0.3**	**76.9±0.5**	93.3±0.5	**95.9±0.8**	91.2±0.8

5.4 Comparison with OSL without covariate shift

In this section, we consider the general open-set setting, where the marginal distribution of known classes is fixed. In Table 4, we compare with the previous open-set methods. An interesting finding is that our framework still outperforms the SOTA methods without covariate shift. It indicates our framework effectively learns the discriminative representation to semantic shift.

6 Conclusion

In this paper, we study the *open-set learning under covariate shift* problem, to reject the unseen classes and recognize the seen classes under covariate shift. We first find the connection between SCL and performance generalization. We theoretically justify SCL could promote the conditional invariance. Furthermore, we propose a theoretically grounded framework, SC–OSG, to learn invariant representation by synthetic multi-source samples. The effectiveness of the proposed framework is clearly verified on real-world datasets.

Appendix

A Proof of Theorem 2

Proof For similarity, we define $V_1 = P^{raw}(Z|Y), V_2 = P^{aug}(Z|Y)$. In words, the joint distribution $p(V_1, V_2)$ means they derive from the same semantic $y \in Y$. The data augmentation $\mathcal{A}$ does not change the semantic causes. Thus, we have positive pairs (arise from the same label y). The loss in the Eq. 1 is the categorical cross-entropy of classifying the positive pair correctly.

Formally, we consider the joint distribution $p(V_1, V_2)$ and the product $p(V_1)p(V_2)$. Let us define a distribution q with latent variable C which decides whether a tuple $(g(v_1), g(v_2))$ was drawn from the joint ($C = 1$, the same semantic Y) or product of marginals ($C = 0$) :

$$q(V_1, V_2|C = 1) = p(V_1, V_2), q(V_1, V_2|C = 0) = p(V_1)p(V_2) \tag{A 1}$$

Given N_c positive pair (drawn from the joint distribution, i.e., the same label provided to V_1 and V_2) for every N negative pairs (drawn from the product of marginals) independent randomly drawn inputs provided to V_1 and V_2. The priors on the latent C are: $q(C = 1) = \frac{N_c}{N+N_c}, q(C = 0) = \frac{N_c}{N+N_c}$. By Bayes's rule, the posterior for class $C = 1$ is given by:

$$\begin{aligned} q(C = 1|V_1, V_2) &= \frac{q(V_1, V_2|C = 1)q(C = 1)}{q(V_1, V_2|C = 1)q(C = 1) + q(V_1, V_2|C = 0)q(C = 0)} \\ &= \frac{N_c p(V_1, V_2)}{N_c p(V_1, V_2) + Np(V_1)p(V_2)} \end{aligned} \tag{A 2}$$

We could find minimize the supervised contrastive loss (Equation 1) is proportional to maximizing the posterior $q(C = 1|V_1, V_2)$. Moreover, we have:

$$\begin{aligned} \log q(C = 1|V_1, V_2) &= \log \frac{p(V_1, V_2)}{p(V_1, V_2) + N/N_c p(V_1)p(V_2)} \\ &= -\log(1 + N/N_c \frac{p(V_1)p(V_2)}{p(V_1)p(V_2)}) \leq -\log(N/N_c) + \log \frac{p(V_1, V_2)}{p(V_1)p(V_2)} \\ I(V_1, V_2) &\geq \log(N/N_c) + \mathbb{E}_{q(V_1,V_2|C=1)} \log q(C = 1|V_1, V_2) \end{aligned}$$

Thus maximizating $\mathbb{E}_{q(V_1,V_2|C=1)} \log q(C = 1|V_1, V_2)$, increases a lower bound on mutual information $I(V_1, V_2)$. Through the supervised singal Y, we could regrad the $\mathbb{E}_{q(V_1,V_2|C=1)} \log q(C = 1|V_1, V_2)$ as a binary classification problem, i.e., whether the samples x_i, x_j from the same semantic Y. And the Eq. 1 could address it.

For the ideal representation Z which maximizes the $I(V_1;V_2)$, we would have $P(V_1) = P(V_2)$. It means that we align the representation distribution between the raw distribution and augmented distribution.

□

B Open-set classification rate

We split testing data into seen classes D_s and unseen class D_u. When given threshold ϵ, we can define Correct Classification Rate(CCR) and False Positive Rate(FPR) under open set learning.

$$FPR(\epsilon) = |\{x|x \in D_u \, and \max_c P(y = c|x) \geq \epsilon\}|/|D_u|$$

$$CCR(\epsilon) = |\{x|x \in D_k \, and \max_c P(y = c|x) \geq \epsilon \, and \arg\max_c P(y = c|x) = y_{true}\}|/|D_k|$$

Varying the threshold ϵ from small one to large one, we plot CCR versus FPR. OSCR can be computed from the area under the curve.

Author Contributions Conceptualization: J-JS; Methodology: X-WY; Writing—original draft preparation: J-JS, X-WY; Writing—review and editing: L-ZG, J-JS; Funding acquisition: L-ZG

Funding This research was supported by the NSFC (62176118, 61921006).

Data availibility statement Not Applicable.

Declarations

Conflicts of interest The author declares that he has no conflict of interest.

Code availability Not applicable.

Consent to participate Not applicable.

Consent for publication Not applicable.

Ethics approval Not applicable.

References

Ahuja, K., Shanmugam, K., Varshney, K.R., et al. (2020). Invariant risk minimization games. In *Proceedings of the 37th international conference on machine learning* (pp. 145–155).

Arjovsky, M., Bottou, L., Gulrajani, I. et al. (2019). Invariant risk minimization. CoRR abs/1907.02893

Bachman, P., Hjelm, R. D., & Buchwalter, W. (2019). Learning representations by maximizing mutual information across views. *Advances in Neural Information Processing Systems, 32*, 15509–15519.

Baktashmotlagh, M., Faraki, M., Drummond, T., et al. (2019). Learning factorized representations for open-set domain adaptation. In *Proceedings of the 7th international conference on learning representations*

Bendale, A. & Boult, T.E. (2016). Towards open set deep networks. In *IEEE conference on computer vision and pattern recognition* (pp. 1563–1572).

Busto, P.P. & Gall, J. (2017). Open set domain adaptation. In *IEEE international conference on computer vision* (pp. 754–763).

Caputo, B., Müller, H., Martínez-Gómez, J., et al. (2014). Imageclef 2014: Overview and analysis of the results. In: *Information access evaluation. Multilinguality, multimodality, and interaction - 5th international conference of the CLEF initiative* (pp. 192–211).

Chen, G., Peng, P., Wang, X., et al. (2021). Adversarial reciprocal points learning for open set recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*

Chen, T., Kornblith, S., Norouzi, M., et al. (2020). A simple framework for contrastive learning of visual representations. In *Proceedings of the 37th international conference on machine learning* (pp. 1597–1607).

Christie, G.A., Fendley, N., Wilson, J., et al. (2018). Functional map of the world. In *IEEE/CVF conference on computer vision and pattern recognition* (pp. 6172–6180).

Chuang, C., Robinson, J., Lin, Y., et al. (2020). Debiased contrastive learning. *Advances in Neural Information Processing Systems, 33*, 8765–8775.

Creager, E., Jacobsen, J. & Zemel, R.S. (2021). Environment inference for invariant learning. In *Proceedings of the 38th international conference on machine learning* (pp. 2189–2200).

Da, Q., Yu, Y. & Zhou, Z. (2014). Learning with augmented class by exploiting unlabeled data. In *Proceedings of the 28th AAAI conference on artificial intelligence* (pp. 1760–1766).

Deng, J., Dong, W., Socher, R., et al. (2009). Imagenet: A large-scale hierarchical image database. In *IEEE computer society conference on computer vision and pattern recognition* (pp. 248–255).

Dhamija, A. R., Günther, M., & Boult, T. E. (2018). Reducing network Agnostophobia. *Advances in Neural Information Processing Systems, 31,* 9175–9186.

Fang, Z., Lu, J., Liu, A. et al. (2021). Learning bounds for open-set learning. In *Proceedings of the 38th international conference on machine learning* (pp. 3122–3132).

Fei, G. & Liu, B. (2016). Breaking the closed world assumption in text classification. In *The 2016 conference of the North American chapter of the association for computational linguistics: human language technologies* (pp. 506–514).

Ganin, Y. & Lempitsky, V.S. (2015). Unsupervised domain adaptation by backpropagation. In *Proceedings of the 32nd international conference on machine learning* (pp. 1180–1189).

Geng, C., Huang, S., & Chen, S. (2021). Recent advances in open set recognition: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 43*(10), 3614–3631.

HaoChen, J. Z., Wei, C., Gaidon, A., et al. (2021). Provable guarantees for self-supervised deep learning with spectral contrastive loss. *Advances in Neural Information Processing Systems, 34*, 5000–5011.

He, K., Zhang, X., Ren, S. et al. (2016). Deep residual learning for image recognition. In *IEEE conference on computer vision and pattern recognition* (pp. 770–778).

He, Y., Shen, Z., & Cui, P. (2021). Towards non-I.I.D.image classification: A dataset and baselines. *Pattern Recognition, 110,* 107383.

Hendrycks, D. & Gimpel, K. (2017). A baseline for detecting misclassified and out-of-distribution examples in neural networks. In *Proceedings of the 5th international conference on learning representations*.

Hendrycks, D., Mazeika, M., Kadavath, S., et al. (2019). Using self-supervised learning can improve model robustness and uncertainty. *Advances in Neural Information Processing Systems, 32*, 15637–15648.

Hjelm, R.D., Fedorov, A., Lavoie-Marchildon, S. et al. (2019). Learning deep representations by mutual information estimation and maximization. In *Proceedings of the 7th international conference on learning representations*.

Huang, Z., Xue, C., Han, B., et al. (2021). Universal semi-supervised learning. *Advances in Neural Information Processing Systems, 34*, 26714–26725.

Huang, Z., Yang, J. & Gong, C. (2022). They are not completely useless: Towards recycling transferable unlabeled data for class-mismatched semi-supervised learning. *IEEE Transactions on Multimedia*.

Khosla, P., Teterwak, P., Wang, C., et al. (2020). Supervised contrastive learning. *Advances in Neural Information Processing Systems, 33*, 18661–18673.

Kim, D., Yoo, Y., Park, S., et al. (2021). Selfreg: Self-supervised contrastive regularization for domain generalization. In *IEEE/CVF international conference on computer vision* (pp. 9599–9608).

LeCun, Y., Bengio, Y., & Hinton, G. (2015). Deep learning. *Nature, 521*(7553), 436–444.

Liang, S., Li, Y., Srikant, R. (2018). Enhancing the reliability of out-of-distribution image detection in neural networks. In *Proceedings of the 6th international conference on learning representations*.

Liao, H. (2013). Speaker adaptation of context dependent deep neural networks. In *IEEE international conference on acoustics, speech and signal processing* (pp. 7947–7951).

Liu, J., Hu, Z., Cui, P. et al. (2021). Heterogeneous risk minimization. In *Proceedings of the 38th international conference on machine learning* (pp. 6804–6814).

Liu, S., Garrepalli, R., Dietterich, T.G. et al. (2018). Open category detection with PAC guarantees. In *Proceedings of the 35th international conference on machine learning* (pp. 3175–3184).

Lundberg, S. M., & Lee, S. (2017). A unified approach to interpreting model predictions. *Advances in Neural Information Processing Systems, 30*, 4765–4774.

Luo, Y., Wang, Z., Huang, Z., et al. (2020). Progressive graph learning for open-set domain adaptation. In *Proceedings of the 37th international conference on machine learning* (pp. 6468–6478).

Neal, L., Olson, M.L., Fern, X.Z., et al. (2018). Open set learning with counterfactual images. In *Proceedings of the 15th European conference on computer vision* (pp. 620–635).

Rejmanek, D., Hosseini, P. R., Mazet, J. A., et al. (2015). Evolutionary dynamics and global diversity of influenza a virus. *Journal of Virology, 89*(21), 10993–11001.

Ribeiro, M.T., Singh, S., & Guestrin, C. (2016). Why should I trust you?: Explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining* (pp. 1135–1144).

Selvaraju, R. R., Cogswell, M., Das, A., et al. (2020). Grad-cam: Visual explanations from deep networks via gradient-based localization. *International Journal of Computer Vision, 128*(2), 336–359.

Shu, Y., Cao, Z., Wang, C. et al. (2021). Open domain generalization with domain-augmented meta-learning. In *IEEE/CVF conference on computer vision and pattern recognition* (pp. 9624–9633).

Tack, J., Mo, S., Jeong, J., et al. (2020). CSI: Novelty detection via contrastive learning on distributionally shifted instances. *Advances in Neural Information Processing Systems, 33,* 11839–11852.

Tian, Y., Krishnan, D. & Isola, P. (2020a). Contrastive representation distillation. In *Proceedings of the 8th international conference on learning representations*.

Tian, Y., Sun, C., Poole, B., et al. (2020). What makes for good views for contrastive learning? *Advances in Neural Information Processing Systems, 33*, 6827–6839.

Vaze, S., Han, K., Vedaldi, A., et al. (2022). Open-set recognition: A good closed-set classifier is all you need. In *Proceedings of the 10th international conference on learning representations*.

Wang, Y., Li, H. & Kot, A.C. (2020). Heterogeneous domain generalization via domain mixup. In *IEEE international conference on acoustics, speech and signal processing* (pp. 3622–3626).

Wen, Z. & Li, Y. (2021). Toward understanding the feature learning process of self-supervised contrastive learning. In *Proceedings of the 38th international conference on machine learning* (pp. 11,112–11,122).

Winkens, J., Bunel, R., Roy, A.G., et al. (2020). Contrastive training for improved out-of-distribution detection. CoRR abs/2007.05566.

Wong, K., Wang, S., Ren, M., et al. (2019). Identifying unknown instances for autonomous driving. In: Kaelbling LP, Kragic D, Sugiura K (eds) *Proceedings of the 3rd annual conference on robot learning* (pp. 384–393).

Yang, Y., Hou, C., Lang, Y., et al. (2019). Open-set human activity recognition based on micro-Doppler signatures. *Pattern Recognition, 85,* 60–69.

Yoshihashi, R., Shao, W., Kawakami, R., et al. (2019) Classification-reconstruction learning for open-set recognition. In *IEEE/CVF conference on computer vision and pattern recognition* (pp. 4016–4025).

Yu, F., Chen, H., Wang, X., et al. (2020). BDD100K: A diverse driving dataset for heterogeneous multitask learning. In *IEEE/CVF conference on computer vision and pattern recognition* (pp. 2633–2642).

Yu, Q. & Aizawa, K. (2019). Unsupervised out-of-distribution detection by maximum classifier discrepancy. In *IEEE/CVF international conference on computer vision* (pp. 9517–9525).

Yun, S., Han, D., Chun, S., et al. (2019). Cutmix: Regularization strategy to train strong classifiers with localizable features. In *IEEE/CVF international conference on computer vision* (pp. 6022–6031).

Zhang, H., Cissé, M., Dauphin, Y.N. et al. (2018). Mixup: Beyond empirical risk minimization. In *Proceedings of the 6th international conference on learning representations*.

Zhang, X., Cui, P., Xu, R., et al. (2021). Deep stable learning for out-of-distribution generalization. In *IEEE/CVF conference on computer vision and pattern recognition* (pp. 5372–5382).

Zhou, D., Ye, H., & Zhan, D. (2021a). Learning placeholders for open-set recognition. In *IEEE/CVF conference on computer vision and pattern recognition* (pp. 4401–4410).

Zhou, Z., Guo, L., Cheng, Z., et al. (2021). STEP: out-of-distribution detection in the presence of limited in-distribution labeled data. *Advances in Neural Information Processing Systems, 34*, 29168–29180.

Zimmermann, R.S., Sharma, Y., Schneider, S., et al. (2021). Contrastive learning inverts the data generating process. In *Proceedings of the 38th international conference on machine learning* (pp. 12979–12990).

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.