Full-Spectrum Out-of-Distribution Detection

Jingkang Yang · Kaiyang Zhou · Ziwei Liu

Received: 10 April 2022 / Accepted: 15 April 2023 / Published online: 13 June 2023
© The Author(s), under exclusive license to Springer Science+Business Media, LLC, part of Springer Nature 2023

Abstract
Existing out-of-distribution (OOD) detection literature clearly defines semantic shift as a sign of OOD but does not have a consensus over covariate shift. Samples experiencing covariate shift but not semantic shift from the in-distribution (ID) are either excluded from the test set or treated as OOD, which contradicts the primary goal in machine learning—being able to generalize beyond the training distribution. In this paper, we take into account both shift types and introduce full-spectrum OOD (F-OOD) detection, a more realistic problem setting that considers both detecting semantic shift and being tolerant to covariate shift; and design three benchmarks. These new benchmarks have a more fine-grained categorization of distributions (i.e., training ID, covariate-shifted ID, near-OOD, and far-OOD) for the purpose of more comprehensively evaluating the pros and cons of algorithms. To address the F-OOD detection problem, we propose SEM, a simple feature-based semantics score function. SEM is mainly composed of two probability measures: one is based on high-level features containing both semantic and non-semantic information, while the other is based on low-level feature statistics only capturing non-semantic image styles. With a simple combination, the non-semantic part is canceled out, which leaves only semantic information in SEM that can better handle F-OOD detection. Extensive experiments on the three new benchmarks show that SEM significantly outperforms current state-of-the-art methods. Our code and benchmarks are released in https://github.com/Jingkang50/OpenOOD.

Keywords Out-of-distribution detection · AI safety · Model trustworthy

1 Introduction
State-of-the-art deep neural networks are notorious for their overconfident predictions on out-of-distribution (OOD) data (Yang et al., 2021), defined as those not belonging to in-distribution (ID) classes. Such behavior makes real-world deployments of neural network models untrustworthy and could endanger users involved in the systems. To solve the problem, various OOD detection methods have been proposed in the past few years (Hendrycks & Gimpel, 2017; Liang et al., 2017; Liu et al., 2020; Lee et al., 2018; Ren et al., 2019; Van Amersfoort & Smith, 2020; Sastry et al., 2020; Wang et al., 2022). The main idea for an OOD detection algorithm is to assign to each test image a score that can represent the likelihood of whether the image comes from in- or out-of-distribution. Images whose scores fail to pass a threshold are rejected, and the decision-making process should be transferred to humans for better handling.

A critical problem in existing research of OOD detection is that only semantic shift is considered in the detection benchmarks while covariate shift—a type of distribution shift that is mainly concerned with changes in appearances like image contrast, lighting or viewpoint—is either excluded from the evaluation stage or simply treated as a sign of OOD (Yang et al., 2021), which contradicts with the primary goal in machine learning, i.e., to generalize beyond the training distribution (Zhou et al., 2022).

To well reconcile OOD generalization and OOD detection tasks, in this paper, we introduce a more challenging yet realistic variant of OOD detection called full-spectrum out-of-distribution detection, or F-OOD detection. The new setting takes into account both the detection of semantic shift and the ability to generalize, i.e., recognize covariate-shifted data as ID. To this end, we design three
benchmarks, namely DIGITS, OBJECTS and COVID, each targeting a specific visual recognition task and together constituting a comprehensive testbed. We also provide a more fine-grained categorization of distributions for the purpose of thoroughly evaluating an algorithm. Specifically, we divide distributions into four groups: training ID, covariate-shifted ID, near-OOD, and far-OOD (the latter two are inspired by a recent study (Ming et al., 2021)). Figure 1a shows example images from the DIGITS benchmark: the covariate-shifted images contain the same semantics as the training images, i.e., digits from 0 to 9, and should be classified as ID, whereas the two OOD groups clearly differ in semantics but represent two different levels of covariate shift.

Ideally, an OOD detection system is expected to produce high scores for samples from the training ID and covariate-shifted ID groups, while assigning low scores to samples from the two OOD groups. However, when applying a state-of-the-art OOD detection method, e.g., the energy-based EBO (Liu et al., 2020), to the proposed benchmarks like DIGITS (see Fig. 1b), we observe that the resulting scores can not solve the setting where covariance shift samples are introduced into the in-distribution. As shown in Fig. 1b, all data are classified as ID including both near-OOD and far-OOD samples.

To address the more challenging but realistic F-OOD detection problem, we propose SEM, a simple feature-based SEMantics score function. Unlike existing score functions that are based on either marginal distribution (Liu et al., 2020) or predictive confidence (Hendrycks & Gimpel, 2017), SEM leverages feature from both top and shallow layers to deduce a single score that is only relevant to semantics, hence more suitable for identifying semantic shift while ensuring robustness under covariate shift. Specifically, SEM is mainly composed of two probability measures: one is based on high-level features containing both semantic and non-semantic information, while the other is based on low-level feature statistics only capturing non-semantic image styles. With a simple combination, the non-semantic part is canceled out, which leaves only semantic information in SEM. Figure 1c illustrates that SEM’s scores are much clearer to distinguish between ID and OOD.

We summarize the contributions of this paper as follows. (1) For the first time, we introduce the full-spectrum OOD detection problem, which represents a more realistic scenario considering both semantic and covariate shift in the evaluation pipeline. (2) Three benchmark datasets are designed for research on F-OOD detection. They cover a diverse set of recognition tasks and have a detailed categorization over distributions. (3) A simple yet effective OOD detection score function called SEM is proposed. Through extensive experiments on the three new benchmarks, we demonstrate that SEM significantly outperforms current state-of-the-art methods in F-OOD detection. The source code and new datasets are open-sourced in https://github.com/Jingkang50/OpenOOD.

2 Related Work

Out-of-distribution detection, or OOD detection, aims to detect test samples that are drawn from a distribution that is different from the training distribution (Yang et al., 2021). The definition of OOD samples can be different under different scenarios. For most machine learning tasks, the distribution should refer to “label distribution”, which means that OOD samples should not have overlapping labels w.r.t. training data. On the other hand, for applications where models are by-design nontransferable to other domains, such as many deep reinforcement learning tasks like game AI (Vinyals et al., 2017; Andreas et al., 2019), models are expected to be conservative and only make predictions for samples exactly from the training distribution, so that the samples with covariate shift shall be considered as OOD. Nevertheless, detecting semantic shift is still the mainstream of OOD detection, and does not conflict with the goal of OOD generalization.
The key idea of the OOD detection method is to design a metric, known as score function, to assess whether a test sample comes from in- or out-of-distribution. The most commonly used metric is based on the conditional probability \( p(y|x) \). An early OOD detection method is maximum soft-max probability (MSP) (Hendrycks & Gimpel, 2017), which is motivated by the observation that deep neural networks tend to give lower confidence to misclassified or OOD data. A follow-up work ODIN (Liang et al., 2017) applies a temperature scaling parameter to soften the probability distribution and further improves the performance by injecting adversarial perturbations to the input. Model ensembling has also been found effective in enhancing robustness in OOD detection (Hyunsun et al., 2018; Vyas et al., 2018). Recent work ViM (Wang et al., 2022) integrates both the norm of feature residuals against the principal space formed by training features and the original logits to compute the score for OOD-ness.

Another direction is to design the metric in a way that reflects the marginal probability \( p(x) \). Liu et al. (2020) connect their OOD score to the marginal distribution using an energy-based formulation, which essentially sums up the prediction logits over all classes. Lee et al. (2018) assume the source data follow a normal distribution and learn a Mahalanobis distance to compute the discrepancy between test images and the estimated distribution parameters. Generative modeling has also been investigated to estimate a likelihood ratio for scoring test images (Hyunsun et al., 2018; Ren et al., 2019), Serrà et al., (2020).

Some methods exploit external OOD datasets. For example, Hendrycks et al. (2019) extend MSP by training the model to produce uniform distributions on external OOD data. Later works introduce re-sampling strategy (Li & Vasconcelos, 2020) and cluster-based methodology (Yang et al., 2021) to better leverage the background data. However, this work does not use external OOD datasets for model design.

Different from all existing methods, our approach aims to address a more challenging scenario, i.e., OOD detection, F-OOD detection, which has not been investigated in the literature but is critical to real-world applications. The experiments show that current state-of-the-art methods mostly fail in the new setting while our approach gains significant improvements.

**3 Methodology**

**3.1 Feature-Based Semantics Score Function**

Key to detect out-of-distribution (OOD) data lies in the design of a score function, which is used as a quantitative measure to distinguish between in- and out-of-distribution data. Our idea is to design the function in such a way that the degree of semantic shift is effectively captured, i.e., the designed score is only sensitive to semantic shift while being robust to covariate shift. For data belonging to the in-distribution classes, the score is high, and vice versa.

**3.1.1 Formulation**

Our score function, called SEM, has the following design:

\[
\text{SEM}(x) = \log p(x_s),
\]

where \( x \) denotes image features learned by a neural network; and \( x_s \) denotes features that only capture the semantics. The probability \( p(x_s) \) can be computed by a probabilistic model, such as a Gaussian mixture model.

The straightforward way to model \( x_s \) is to learn a neural network for image recognition and hope that the output features \( x \) only contain semantic information, i.e., \( \text{SEM}(x) = \log p(x) \).

However, numerous studies have suggested that the output features \( x \) often contain both semantic and non-semantic information while decoupling them is still an open research problem (Zhou et al., 2022; Chuang et al., 2021; Peng et al., 2019). Let \( x_n \) denote non-semantic features, we assume that semantic features \( x_s \) and non-semantic features \( x_n \) are generated independently, namely

\[
p(x) = p(x_s)p(x_n).
\]

We propose a simple method to model the score function so that it becomes only relevant to the semantics of an image. This is achieved by leveraging low-level feature statistics, i.e., means and standard deviations, learned in a CNN, which have been shown effective in capturing image styles that are essentially irrelevant to semantics (Zhou et al., 2021). Specifically, the score function in Eq. (1) is rewritten as

\[
\text{SEM}(x) = \log p(x_s) = \log \frac{p(x_s)p(x_n)}{p(x_n)} = \log \frac{p(x)}{p(x_n)}.
\]

where \( p(x) \) is computed using the output features while \( p(x_n) \) is based on low-level feature statistics. It is noticed that the SEM score can also be referred to as likelihood ratio (Ren et al., 2019). More discussion on the similarity and differences with the work will be described at the end of Sect. 3.1.

Below we first discuss how to compute feature statistics and then detail the approach of how to model the distributions for \( x \) and \( x_n \).
Fig. 2 Overview of our Methodology. a The computation of SEM score function for OOD detection. SEM combines the estimation of \( p(x) \) (using top-layer features to capture both semantic and non-semantic information) and \( p(x_n) \) (using low-level feature statistics to only capture non-semantic information) with Eq. (3) for better concentration on semantics. b The fine-tuning scheme to enhance source-awareness for better estimating \( p(x_n) \). The main idea is to pull together the instance-level CNN feature statistics of in-distribution data to make them more compact, while pushing away those of synthetic OOD data, which are obtained by negative data augmentation such as Mixup (Zhang et al., 2018).

3.1.2 Feature Statistics Computation

Instance-level feature statistics have been widely used in the style transfer community for manipulating image style (Huang & Belongie, 2017). Given a set of CNN feature maps \( z \in \mathbb{R}^{C \times H \times W} \) with \( C, H \) and \( W \) denoting the number of channels, height, and width, their feature statistics, \( \mu \in \mathbb{R}^C \) and standard deviations \( \sigma \in \mathbb{R}^C \), are computed across the spatial dimension within each channel \( c = \{1, 2, \ldots, C\} \),

\[
\mu_c = \frac{1}{HW} \sum_{h=1}^{H} \sum_{w=1}^{W} z_{c,h,w},
\]

\[
\sigma_c = \left( \frac{1}{HW} \sum_{h=1}^{H} \sum_{w=1}^{W} (z_{c,h,w} - \mu_c)^2 \right)^{1/2}.
\]

As shown in Zhou et al. (2021), the feature statistics in shallow CNN layers are strongly correlated with domain information (i.e., token@token@noet, image style) while those in higher layers pick up more semantics. Therefore, we choose to extract feature statistics in the first CNN layer and represent \( x_n \) by concatenating the means and standard deviations, \( x_n = [\mu, \sigma]^T \).

3.1.3 Distribution Modeling

For simplicity, we model \( p(x) \) and \( p(x_n) \) in Eq. (3) using the same approach, which consists of two steps: dimension reduction and distribution modeling. Below we only discuss \( p(x) \) for clarity.

Motivated by the manifold assumption in Bengio et al. (2013) that suggests data typically lie in a manifold of much lower dimension than the input space, we transform features \( x \) to a new low-dimensional space, with a hope that the structure makes it easier to distinguish between in- and out-of-distribution. To this end, we propose a variant of the principal component analysis (PCA) approach. Specifically, rather than maximizing the variance for the entire population, we maximize the sum of variances computed within each class with respect to the transformation matrix. In doing so, we can identify a space that is less correlated with classes.

Given a training dataset, we build a Gaussian mixture model (GMM) to capture \( p(x) \). Formally, \( p(x) \) is defined as

\[
p(x) = \sum_{m=1}^{M} \lambda_m N(\alpha_m, \beta_m),
\]

where \( M \) denotes the number of mixture components, \( \lambda_m \) the mixture weight s.t. \( \sum_{m=1}^{M} \lambda_m = 1 \), and \( \alpha_m \) and \( \beta_m \) the means and variances of a normal distribution. A GMM model can be efficiently trained by the expectation-maximization (EM) algorithm.

3.1.4 Comparison with Likelihood-Ratio OOD Detection Ren et al. (2019)

While both methods use the likelihood-ratio score function to remove the background/covariate-shift scores while emphasizing semantics, the main difference is in the way we model \( p(x) \) and \( p(x_n) \). In Ren et al. (2019), two generative PixelCNN++ models are used to model \( p(x) \) and \( p(x_n) \) separately.
Moreover, to model \( p(x_n) \), Ren et al. (2019) perturbs inputs in the pixel space by randomly flipping input pixel values to one of the 256 possible values with a certain mutation rate. However, in our proposed method, we use a GMM to estimate \( p(x) \) and \( p(x_n) \) using simply-processed features from a standard classification network. This makes our method much simpler and computationally more feasible than Ren et al. (2019), as SEM score can be considered as an add-on for any pre-trained classifier, while (Ren et al., 2019) requires the use of generative models such as PixelCNN++, making it difficult to conduct fair comparisons with other classifier-based OOD detectors like MSP (Hendrycks & Gimpel, 2017).

### 3.2 Source-Awareness Enhancement

While feature statistics exhibit a higher correlation with source distributions (Zhou et al., 2021), the boundary between in- and out-of-distribution in complicated real-world data is not guaranteed to be clear enough for differentiation. Inspired by Liu et al. (2020) who fine-tune a pre-trained model to increase the energy values assigned to OOD data and lower down those for ID data, we propose a fine-tuning scheme to enhance source awareness in feature statistics. An overview of the fine-tuning scheme is illustrated in Fig. 2b.

#### 3.2.1 Negative Data Augmentation

The motivation behind our fine-tuning scheme is to obtain a better estimate of the non-semantic score, in hope that it will help SEM better capture the semantics with the combination in Eq. (3). This can be achieved by explicitly training feature statistics of ID data to become more compact while pushing OOD data’s feature statistics away from the ID support areas. A straightforward way is to collect auxiliary OOD data like Liu et al. (2020) for building a contrastive objective. In this work, we propose a more efficient way by using negative data augmentation (Sinha et al., 2021) to synthesize OOD samples. The key idea is to choose data augmentation methods to easily generate samples with covariate shift. One example of augmentation is Mixup (Zhang et al., 2018).

#### 3.2.2 Learning Objectives

Given a source dataset \( S = \{ (x, y) \} \), we employ negative data augmentation methods \( \text{aug}(\cdot) \) to synthesize an OOD dataset \( S_{\text{aug}} = \{ (x', y) \} \) where \( x' = \text{aug}(x) \). For fine-tuning, we combine a classification loss \( L_{\text{cls}} \) with a source-awareness enhancement loss \( L_{\text{src}} \). These two losses are formally defined as

\[
L_{\text{cls}} = - \sum_{(x, y) \sim S} \log p(y|x),
\]

and

\[
L_{\text{src}} = \sum_{x' \sim S_{\text{aug}}} p(x') - \sum_{x \sim S} p(x),
\]

where the marginal probability \( p(x) \) is computed based on a GMM model described previously. Note that the GMM model is updated every epoch to adapt to the changing features.

After fine-tuning, we learn a new GMM model using the original source dataset. This model is then used to estimate the marginal probability \( p(x) \) at test time.

### 4 F-OOD Benchmarks

To evaluate full-spectrum out-of-distribution (F-OOD) detection algorithms, we design three benchmarks: DIGITS, OBJECTS, and COVID. Examples of DIGITS are shown in Fig. 1 and the other two are shown in Fig. 3. A numerical summary of the benchmarks is shown in Table 2.

#### 4.1 Benchmark-1: DIGITS

We construct the DIGITS benchmark based on the popular digit datasets: MNIST (LeCun et al., 1998), which contains 60,000 images for training. During testing, the model will be exposed to 10,000 MNIST test images, with 26,032 covariate-shifted ID images from SVHN (Netzer et al., 2011) and another 9,298 from USPS (Hull, 1994). The near-OOD datasets are notMNIST (Yaroslav, 2011) and FashionMNIST (Xiao et al., 2017), which share a similar background style with MNIST. The far-OOD datasets consist of a textural dataset (Texture (Cimpoi et al., 2014)), two object datasets (CIFAR-10 (Krizhevsky et al., 2009) & Tiny-ImageNet (Russakovsky et al., 2015)), and one scene dataset (Places365 (Zhou et al., 2017)). The CIFAR-10 and Tiny-ImageNet test sets have 10,000 images for each. The Places365 test set contains 36,500 scene images.

#### 4.2 Benchmark-2: OBJECTS

The OBJECTS benchmark is built on top of CIFAR-10 (Krizhevsky et al., 2009), which contains 50,000 images for training. During testing, the model will be exposed to 10,000 CIFAR-10 test images, and another 10,000 images selected from ImageNet-22K (Russakovsky et al., 2015) with
Fig. 3 Examples for the two F-OOD detection benchmarks: COVID and OBJECTS. Each benchmark consists of a training ID dataset, two covariate-shifted ID datasets, two near-OOD datasets, and four far-OOD datasets.

4.3 Benchmark-3: COVID

We construct a real-world benchmark to show the practical value of F-OOD. We simulate the scenario where an AI-assisted diagnostic system is trained to identify COVID-19 infection from chest X-ray images. The training data come from a single source (e.g., a hospital) while the covariate-shifted ID test data are from other hospitals or machines, which the system needs to be robust and produce reliable predictions. Specifically, we refer to the COVID-19 chest X-ray dataset review (Santa Cruz et al., 2021), and use the large-scale image collection from Valencian Region Medical ImageBank Vayá et al., (2020) (referred to as BIMCV) as training ID images (randomly sampled 2443 positive cases and 2501 negative cases with necessary cleaning). Images from two other sources, i.e., ActMed with 132 positive images), and Hannover (Winther et al., 2020) (from Hannover Medical School with 243 positive images), are considered as the covariate-shifted ID group. OOD images are from completely different classes. Near-OOD images are obtained from other medical datasets, i.e., the RSNA Bone Age dataset with 200 bone X-ray images (RSNA, 2017) and 544 COVID CT images (Yang et al., 2020). Far-OOD samples are defined as those with drastic visual and concept differences from the ID images. We use MNIST, CIFAR-10, Texture and Tiny-ImageNet.

4.4 Evaluation Metrics

In the F-OOD setting, different datasets belonging to one OOD type (i.e., covariate-shifted ID or training ID) are grouped together. We also report the performance on contrasting covariate-shifted ID with training ID, although covariate-shifted ID are not OOD samples. We use three metrics to evaluate the OOD detection performance, which is detailed as follows: (1) FPR95 stands for false positive rate measured when true positive rate (TPR) sits at 95%. Intuitively, FPR95 measures the portion of samples that are falsely recognized as in-distribution data when most true in-distribution samples are recalled. (2) AUROC refers to the Area Under the Receiver Operating Characteristic curve, which is concerned with both FPR and TPR. (3) AUPR means the Area Under the Precision-Recall curve, which considers both precision and recall. For FPR95, the lower the value, the better the model. For AUROC and AUPR, the higher the value, the better the model.

5 Experiments

5.1 Implementation Details

We conduct experiments on the three proposed F-OOD benchmarks, i.e., DIGITS, OBJECTS, and COVID. In terms of architectures, we use LeNet-5 (LeCun et al., 2015) for DIGITS and ResNet-18 (He et al., 2016) for both OBJECTS and COVID. All models are trained by the SGD optimizer with a weight decay of $5 \times 10^{-4}$ and a momentum of 0.9. For DIGITS and OBJECTS, we set the initial learning rate to 0.1, which is decayed by the cosine annealing rule, and the total epochs to 100. For COVID benchmark, the initial learning rate is set to 0.001 and the model is trained for 200 epochs. When fine-tuning for source-awareness enhancement, the learning rate is set to 0.005 and the total number of epochs is 10. The batch size is set to 128 for all benchmarks.

Notice that the baseline implementations of ODIN (Liang et al., 2017) and MDS (Lee et al., 2018) require a validation set for hyperparameter tuning, we spare a certain...
Table 1  Selected ImageNet-22K classes for OBJECTS benchmark

| Airplane             | Automobile             | Bird                    | Cat                    |
|----------------------|------------------------|-------------------------|------------------------|
| n03365231 floatplane | n04516354 used car     | n01530616 bird          | n02121808 domestic cat |
| n02691156 airplane   | n04285008 sports car   | n01812337 dove          | n02123159 tiger cat    |
| n04552348 warplane   | n02958343 car          | n01562265 robin         | n02122878 tabby        |
| n02686568 aircraft   | n03594945 jeep         | n01539573 sparrow       | n02123394 Persian cat  |
| n02690373 airliner   | n02930766 cab          | n01558594 blackbird     | n02123597 Siamese cat  |

| Deer                 | Dog                    | Frog                   |
|----------------------|------------------------|------------------------|
| n02430045 deer       | n02116738 African hunting dog | n01639765 frog          |
| n02431122 red deer   | n02087122 hunting dog  | n01641577 bullfrog      |
| n02432511 mule deer  | n02105855 Shetland sheepdog | n01644373 tree frog     |
| n02433318 fallow deer| n02109961 Eskimo dog   | n01640846 true frog     |
| n02431976 Japanese deer | n02099601 golden retriever | n01642539 grass frog   |

| Horse                | Ship                   | Truck                  |
|----------------------|------------------------|------------------------|
| n02387254 farm horse | n02965300 cargo ship   | n04490091 truck        |
| n02381460 wild horse | n04194289 ship         | n03417042 garbage truck|
| n02374451 horse      | n03095699 container ship | n03173929 delivery truck|
| n02382948 racehorse  | n02981792 catamaran     | n04467665 trailer truck|
| n02379183 quarter horse | n03344393 fireboat    | n03345487 fire engine  |

We manually find 5 ImageNet-22K classes that belong to each CIFAR-10 class, and pick the first 1000 images from every selected class for the OBJECTS benchmark. A string such as ‘n03365231’ is the synset id for downloading the corresponding class from ImageNet API.

5.2 Results on F-OOD Setting

We first discuss the results on near- and far-OOD datasets. Table 3 summarizes the results where the proposed SEM is compared with current state-of-the-art methods including MSP (Hendrycks & Gimpel, 2017), Mahalanobis distance score (MDS), Energy-based OOD (Liu et al., 2020), and ViM (Wang et al., 2022).

5.2.1 DIGITS Benchmark

For the DIGITS benchmark, SEM gains significant improvements in all metrics (FPR95, AUROC, and AUPR). A huge gain is observed on notMNIST, which is a challenging dataset due to its closeness in the background to the training ID MNIST. While none of the previous softmax/logits-based methods (e.g., glet@tokeneddot, MSP and EBO) are capable to solve the notMNIST problem, the proposed SEM largely reduces the FPR95 metric from 99% to 10.93%, and the AUROC is increased from around 30% to beyond 95%. One explanation of the clear advantage is that the previous output-based OOD detection methods largely depend on the covariate shift to detect OOD samples, while the feature-based MDS (partly relies on top-layer semantic-aware features) and the proposed SEM uses more semantic information, which is critical to distinguish MNIST and notMNIST. In other words, in the MNIST/notMNIST scenario where ID and OOD have high visual similarity, a large dependency on covariate shift while ignorance of the semantic information will lead to the failure of OOD separation. Similar advantages are also achieved with the other near-OOD dataset.

5.2.2 OBJECTS Benchmark

Similar to the DIGITS benchmark, the proposed SEM surpasses the previous state-of-the-art methods on the near-OOD scenario of the OBJECTS benchmark, especially on the more robust metrics of AUROC and AUPR. However, the performance gap is not as large as DIGITS. One explanation is that images in the OBJECTS benchmark are more complex than DIGITS, leading the neural networks to be more semantics-orientated. Therefore, more semantic information is encoded in the previous output-based methods. Neverthe-
Table 2  Summary of F-OOD benchmarks

|                  | DIGITS                  | OBJECTS                  | COVID       |
|------------------|-------------------------|--------------------------|-------------|
| Training ID      | MNIST (60,000 / 10,000) | CIFAR-10 (50,000 / 10,000) | BIMCV (3954 / 891) |
| Covariate-Shifted ID | SVHN: 26,032 USPS: 9298 | ImageNet-10: 9595 CIFAR-10-C: 10,000 | ActMed: 132 Hannover: 243 |
| Near-OOD         | notMNIST: 17,724 FashionMNIST: 10,000 | CIFAR-100: 10,000 TinyImageNet: 8,793 | CT-SCAN: 544 X-Ray-Bone: 200 |
| Far-OOD          | Texture: 5,640 CIFAR-10: 10,000 Tiny-ImageNet: 10,000 Places365: 36,500 | MNIST: 10,000 FashionM- 10,000 Texture: 5640 NIST: 10,000 Textures: 5640 Tiny-ImageNet: 10,000 |

Number of samples for each dataset is reported. Notice that models for F-OOD are only trained from Training ID, which is class-wise (near-)balanced. Training ID reports # training samples / # testing samples.
Table 3  Comparison between previous state-of-the-art methods and the proposed SEM score on F-OOD benchmarks

|                  | FPR95 ↓ | AUROC ↑ | AUPR ↑ |
|------------------|--------|--------|--------|
|                  | MSP    | EBO    | MDS    | ViM   | SEM   | MSP    | EBO    | MDS    | ViM   | SEM   |
|                  |        |        |        |       |       |        |        |        |       |       |
| **DIGITS** (Training ID: MNIST, Covariate-shifted ID: USPS & SVHN) |        |        |        |       |       |        |        |        |       |       |
| notMNIST         | 99.97  | 99.99  | 78.83  | 66.34 | 10.93 | 32.54  | 25.49  | 79.10  | 85.91 | 96.74 |
| FashionMNIST     | 99.90  | 99.98  | 94.68  | 93.78 | 68.63 | 39.71  | 37.64  | 60.42  | 66.45 | 80.20 |
| Mean (Near-OOD)  | 99.93  | 99.98  | 86.75  | 80.06 | 39.78 | 36.12  | 31.56  | 69.76  | 76.18 | 88.47 |
| Texture          | 94.89  | 98.40  | 87.46  | 92.34 | 90.90 | 64.34  | 65.02  | 72.42  | 70.32 | 74.45 |
| CIFAR-10         | 98.01  | 99.62  | 95.47  | 93.60 | 91.57 | 52.22  | 50.95  | 67.96  | 63.23 | 69.20 |
| Tiny-ImageNet    | 97.98  | 99.58  | 96.20  | 96.18 | 93.39 | 52.94  | 51.89  | 64.31  | 62.34 | 67.54 |
| Places365        | 98.68  | 99.65  | 98.06  | 98.02 | 94.15 | 50.22  | 48.95  | 65.42  | 65.32 | 67.63 |
| Mean (Far-OOD)   | 97.39  | 99.31  | 94.30  | 95.04 | 92.50 | 54.93  | 54.20  | 67.53  | 65.30 | 69.73 |
| **OBJECTS** (Training ID: CIFAR-10, Covariate-shifted ID: CIFAR-10-C & ImageNet-10) |        |        |        |       |       |        |        |        |       |       |
| CIFAR-100        | 89.44  | 83.84  | 86.28  | 82.44 | 86.96 | 70.17  | 63.85  | 72.05  | 70.12 | 74.70 |
| Tiny-ImageNet    | 88.22  | 81.58  | 87.45  | 87.32 | 86.59 | 72.92  | 67.97  | 72.91  | 73.24 | 76.76 |
| Mean (Near-OOD)  | 88.83  | 82.71  | 86.87  | 84.88 | 86.77 | 71.55  | 65.91  | 72.50  | 71.68 | 75.73 |
| MNIST            | 93.54  | 92.23  | 84.59  | 90.12 | 99.70 | 69.98  | 54.55  | 77.04  | 73.45 | 75.69 |
| FashionMNIST     | 88.08  | 72.40  | 77.17  | 75.32 | 93.72 | 73.78  | 76.50  | 80.33  | 81.01 | 79.40 |
| Texture          | 85.64  | 75.57  | 72.98  | 70.92 | 82.15 | 74.18  | 68.63  | 72.02  | 73.40 | 79.69 |
| CIFAR-100-C      | 87.26  | 83.64  | 85.53  | 89.32 | 83.92 | 74.12  | 68.37  | 68.13  | 73.32 | 78.89 |
| Mean (Far-OOD)   | 88.63  | 80.96  | 80.07  | 81.42 | 89.87 | 72.27  | 67.01  | 74.38  | 75.30 | 78.42 |
| **COVID** (Training ID: BIMCV, Covariate-shifted ID: ActMed & Hanover) |        |        |        |       |       |        |        |        |       |       |
| CT-SCAN          | 99.80  | 97.35  | 99.39  | 12.56 | 2.24  | 11.31  | 13.14  | 81.21  | 89.23 | 99.51 |
| XRayBone         | 97.00  | 42.00  | 100.00 | 23.35 | 14.50 | 32.08  | 77.80  | 78.72  | 90.37 | 94.97 |
| Mean (Near-OOD)  | 98.40  | 69.67  | 99.69  | 17.96 | 8.37  | 21.70  | 45.47  | 79.96  | 89.80 | 97.24 |
| MNIST            | 98.30  | 0.35   | 100.00 | 1.69  | 24.89 | 99.11  | 80.81  | 98.32  | 100.00| 1.07  |
| CIFAR-10         | 96.32  | 94.67  | 98.02  | 89.61 | 85.58 | 41.12  | 45.23  | 77.05  | 46.30 | 52.50 |
| Texture          | 98.39  | 87.06  | 56.38  | 40.25 | 27.57 | 22.63  | 34.95  | 89.84  | 82.80 | 90.94 |
| Tiny-ImageNet    | 97.78  | 92.73  | 92.11  | 73.96 | 44.99 | 30.26  | 32.69  | 81.99  | 79.32 | 83.42 |
| Mean (Far-OOD)   | 97.70  | 68.70  | 86.63  | 51.38 | 39.54 | 29.73  | 53.20  | 82.42  | 76.69 | 81.72 |

The proposed SEM obtains a consistently better performance on most of the metrics than MSP (Hendrycks & Gimpel, 2017), Energy-based OOD (EBO) score (Liu et al., 2020), Mahalanobis Distance Score (MDS) (Lee et al., 2018), and ViM (Wang et al., 2022), especially on the near-OOD scenarios.

Bold indicates optimal performance at the corresponding setting.
### Table 4
Comparison between previous state-of-the-art methods, the proposed SEM score, and the low-level probabilistic component $p(x_n)$ on classic OOD benchmarks, without the existence of covariate-shifted ID set

|                | FPR95 ↓ | AUROC ↑ | AUPR ↑ |
|----------------|---------|---------|--------|
|                | MSP     | MDS     | EBO    | ViM   | SEM   | $p(x_n)$ | MSP     | MDS     | EBO    | ViM   | SEM   |
| - DIGITS (ID: MNIST) |         |         |        |       |       |          |         |         |        |       |       |
| notMNIST       | 43.09   | 44.06   | 1.77   | 2.64  | 0.78  | 88.77    | 88.44   | 99.67   | 98.76  | 99.50  | 99.79 |
| FashionMNIST   | 2.54    | 1.05    | 0.27   | 29.68 | 40.09 | 0.00     | 99.44   | 99.72   | 99.90  | 88.54  | 95.02 |
| Mean (Near-OOD)| 22.82   | 22.56   | 21.37  | 0.39  | 94.11 | 94.08    | 99.79   | 93.65   | 97.26  | 99.87  |
| Texture        | 2.43    | 0.67    | 0.23   | 2.22  | 90.69 | 0.02     | 99.34   | 99.81   | 99.93  | 99.18  | 77.26 |
| CIFAR-10       | 7.05    | 3.18    | 0.18   | 54.43 | 0.00  | 98.68    | 99.30   | 99.88   | 99.94  | 91.10  | 99.97 |
| Tiny-ImageNet  | 6.28    | 3.13    | 0.55   | 0.89  | 99.52 | 0.00     | 99.87   | 99.37   | 99.79  | 99.69  | 93.70 |
| Places365      | 9.92    | 4.12    | 0.45   | 1.16  | 58.07 | 0.00     | 98.19   | 99.17   | 99.81  | 99.65  | 93.82 |
| Mean (Far-OOD) | 6.42    | 2.78    | 0.35   | 1.12  | 65.68 | 0.01     | 98.75   | 99.41   | 99.85  | 99.62  | 94.97 |
| - OBJECTS (ID: CIFAR-10) |         |         |        |       |       |          |         |         |        |       |       |
| CIFAR-100      | 62.01   | 81.63   | 51.46  | 62.80 | 52.16 | 90.12    | 87.11   | 66.30   | 86.15  | 81.30  | 91.99 |
| Tiny-ImageNet  | 60.69   | 83.76   | 45.02  | 61.54 | 49.19 | 76.24    | 86.62   | 66.79   | 88.58  | 85.74  | 96.11 |
| Mean (Near-OOD)| 61.35   | 82.70   | 48.24  | 62.17 | 50.68 | 83.18    | 86.87   | 66.55   | 87.37  | 85.75  | 96.11 |
| MNIST          | 58.59   | 0.00    | 44.50  | 58.16 | 37.27 | 0.00     | 89.91   | 99.52   | 90.59  | 90.32  | 99.84 |
| FashionMNIST   | 51.87   | 19.69   | 44.94  | 18.22 | 42.38 | 0.00     | 90.88   | 95.78   | 88.39  | 96.78  | 88.96 |
| Texture        | 59.89   | 19.61   | 48.32  | 26.35 | 43.07 | 0.00     | 88.72   | 95.42   | 86.85  | 95.28  | 88.49 |
| CIFAR-100-C    | 57.64   | 48.84   | 41.88  | 53.85 | 39.05 | 44.12    | 89.03   | 64.40   | 89.60  | 89.46  | 90.94 |
| Mean (Far-OOD) | 57.00   | 34.04   | 44.91  | 39.15 | 40.44 | 13.04    | 89.64   | 88.78   | 88.86  | 92.96  | 93.44 |

The previous methods of MSP (Hendrycks & Gimpel, 2017), EBO score (Liu et al., 2020), MDS (Lee et al., 2018), and ViM (Wang et al., 2022), reach good results on the classic benchmark. However, the value of $p(x_n)$ can exceed all the previous methods and achieve a near-perfect result across all the metrics, showing that only taking covariate shift score can completely solve the classic OOD detection benchmark, which, in fact, contradicts the goal of OOD detection. This phenomenon also advocates the significance of the proposed F-OOD benchmark.

Bold indicates optimal performance at the corresponding setting.
less, the proposed SEM method still outperforms others on most of the metrics. We also notice that the SEM score does not reach the best performance on MNIST and FashionMNIST. One explanation is that two black-and-white images in these two datasets inherently contain significant covariate shifts compared to both training ID and covariate-shifted ID, so the scores that are efficient on covariate shift detection can also achieve good results on these datasets. However, these methods fail in near-OOD scenarios, as they might believe CIFAR-10-C should be more likely to be OOD than CIFAR-100.

5.2.3 COVID Benchmark

In this new and real-world application of OOD detection, the proposed SEM score achieves an extraordinary performance on all metrics, which surpasses the previous state-of-the-art methods by a large margin in both near and far-OOD scenarios. The result also indicates that previous output-based methods generally break down on this setting, e.g., their FPR@95 scores are generally beyond 90% in a near-OOD setting which means ID and OOD are totally mixed. However, the proposed SEM achieves around 10% in the near-OOD setting. On far-OOD samples, the output-based methods are still unable to be sensitive to the ID/OOD discrepancy. The phenomenon matches the performance in the DIGITS dataset, where the training data is simple and the logits might learn much non-semantic knowledge to be canceled out.

5.2.4 Observation Summary

We summarize the following two take-away messages from the experiments on all three F-OOD benchmarks: 1) SEM score performs consistently well on near-OOD, which classic output-based methods (e.g., MSP, EBO) majorly fail on. The reason can be that output-based methods use too much covariate shift information for OOD detection, which by nature cannot distinguish between covariate-shifted ID and near-OOD. The proposed SEM score also outperforms the similar feature-based baseline MDS. 2) SEM score sometimes underperforms on far-OOD, with a similar reason that classic OOD detectors use covariate shift to distinguish ID and OOD, which is sometimes sufficient to detect far-OOD samples. Nevertheless, SEM reaches more balanced good results on near-OOD and far-OOD.

5.3 Results on Classic OOD Detection Setting

Table 4 shows the performance on the classic OOD detection benchmark. The result shows that without the introduction of covariate-shifted ID data, the previous methods reach a near-perfect performance on the classic benchmark, which matches the reported results in their origin papers. However, by comparing with Table 3, their performance significantly breakdown when covariate-shifted ID is introduced, showing the fragility of previous methods, and therefore we advocate the more realistic F-OOD benchmark. Furthermore, we also report the results that by using the value of \( p(x_n) \), the score from low-layer feature statistics for detecting covariate shift is shown surprisingly effective on classic OOD benchmark, which exceeds all the previous methods and achieves a near-perfect result across all the metrics. This phenomenon shows that only taking covariate shift scores can completely solve the classic OOD detection benchmark with MNIST, which, in fact, contradicts the goal of OOD detection. It also advocates the significance of the proposed F-OOD benchmark.

5.4 Ablation Study

In this section, we validate the effectiveness of the main components that contribute to the proposed SEM score and also analyze the effects of fine-tuning scheme for source-awareness enhancement. All the experiments in this part are conducted on the DIGITS benchmark.

![T-SNE visualization on DIGITS. It suggests that low-layer feature statistics capture non-semantic information, and top-layer features capture both semantic and non-semantic information](image)

Table 5 Ablation study on the SEM components

| #  | \( p(x_n) \) | \( p(x) \) | NearOOD | FarOOD |
|----|-------------|-------------|---------|--------|
| 1  | ✓           | 87.28       | 60.80   |        |
| 2  | ✓           | 87.28       | 60.80   |        |
| 3  | ✓           | 51.81       | 51.81   |        |
| 4  | ✓           | 86.54       | 61.26   |        |
| 5  | ✓           | 70.27       | 72.58   |        |
| 6  | ✓           | 88.47       | 69.73   |        |
| 7  | ✓           | 88.47       | 69.73   |        |

AUROC is reported for performance evaluation. Several options can be applied to estimate \( p(x_n) \) and \( p(x) \) in Eq. (3). FS denotes the usage of feature statistics, and FF denotes flattened features. T and L mean top-/low-layer feature. e.g., T-FS means top-layer feature statistics. The results show the effectiveness of our SEM score Bold indicates optimal performance at the corresponding setting.
Table 6  Ablation study on the fine-tuning scheme for source-awareness enhancement

| #  | \( L_{cls} \) | \( L_{src}(x) \) | \( L_{src}(x') \) | NearOOD | FarOOD |
|----|-------------|----------------|----------------|---------|--------|
| 1  | 83.03       | 56.65          |              |         |        |
| 2  | ✓           | 86.55          | 64.61         |         |        |
| 3  | ✓           | ✓              | 87.42         | 68.40   |        |
| 4  | ✓           | ✓              | ✓             | 87.27   | 67.92  |
| 5  | ✓           | ✓              | ✓             | 88.47   | 69.73  |

AUROC is reported for performance evaluation. #1 reports the performance before fine-tuning. \( L_{src}(x) \) means fine-tuning without negative augmented data. \( L_{src}(x') \) means only data with negative augmentation is used. The results show the effectiveness of each training loss. Bold indicates optimal performance at the corresponding setting.

5.4.1 Components of SEM

According to Eq. (2) in the Sect. 3, SEM score can be decomposed by the estimations of \( p(x) \) and \( p(x_n) \). While our final SEM score uses output flattened features of the CNN model for \( p(x) \) estimation and low-layer feature statistics for \( p(x_n) \), there are actually several options for the estimation, which is discussed in Table 5. In this analysis, we set top flattened features as the default usage for \( p(x) \) and only explore \( p(x_n) \), which is the key part of SEM score.

Exp#1 shows the result that only uses \( p(x) \) as the final score, which can be interpreted as a simple method using GMM to estimate ID likelihood on the fine-tuning features. Compared to the MDS result in Table 3, this simple method already obtains a better performance on near-OOD. Notice that we use LeNet-5 on DIGITS, the final-layer features are identical to their feature statistics (ref. Exp#2). Therefore, everything is canceled out if \( p(x_n) \) is top-layer feature statistics (ref. Exp#3).

Exp#4 and Exp#6 show a comparison between using low-layer flattened features (L-FF) and low-layer feature statistics (L-FS) only. The performance on detecting covariate-shifted ID shows that both L-FF and L-FS have significant sensitivity to covariate shifts but with poor performance on F-OOD detection. The result indicates that with only the usage of low-level features, the score has a strong correlation to covariate shift but barely contains semantic information, and the feature statistics show stronger characteristics compared to flattened features. This observation indicates our selection of low-level feature statistics for estimating \( p(x_n) \), which is further supported by the results of Exp#5 and Exp#7, and visually illustrated by Fig. 4.

5.4.2 Fine-Tuning Scheme

Here we evaluate the designed fine-tuning scheme of SEM. As elaborated in Sect. 3.2, this learning procedure is designed to enhance the source-aware compactness. Specifically, a source-awareness enhancement loss \( L_{src} \) is proposed to aggregate the ID training data and separate it from the generated negative augmented images at the same time. Table 6 demonstrates the effectiveness of the fine-tuning scheme. When combining both in-distribution training and negative augmented data training, our framework achieves the best performance.

5.4.3 Hyperparameter of M

Table 7 shows the analysis of hyperparameter \( M \). In the DIGITS dataset, \( M = 3 \) leads to a slightly better performance compared to other choices. Nevertheless, the overall difference among various \( M \) is not obvious on near-OOD, showing that the model is robust to the hyperparameter.

5.4.4 Effectiveness with Other Backbones

Table 8 shows the results on the OBJECTS benchmark with another architecture DenseNet-100 (Huang et al., 2017). Compared to the results in the Table 3 using ResNet-18, the result from the larger architecture shows a similar pattern as the proposed SEM score outperforms others in most cases on AUROC and AUPR.

5.4.5 Comparing Outlier Exposure with NDA

Considering that the SEM score leverages synthetic OOD samples in the fine-tuning phase, we also compare the proposed method and outlier exposure (Hendrycks et al., 2019) with negative data augmentation (NDA). Table 9 shows that the SEM score outperforms the OE+NDA baseline to a large extent. The reason is that synthesizing NDA samples as outlier exposure during training can help the model better learn the distinguishment on covariate shift/domain, which contradicts the requirement of F-OOD benchmarks, where the model should be robust to covariate shift.

Table 7  Hyperparameter Selection of the Number of GMM Components K

|   | \( p(x_n) \) M=1 | \( p(x_n) \) M=3 | \( p(x_n) \) M=10 | \( p(x_n) \) M=20 | NearOOD | FarOOD |
|---|-----------------|-----------------|-----------------|-----------------|---------|--------|
| 1 | ✓               | ✓               | ✓               | ✓               | 86.24   | 64.94  |
| 2 | ✓               | ✓               | ✓               | ✓               | 85.81   | 60.17  |
| 3 | ✓               | ✓               | ✓               | ✓               | 84.47   | 63.61  |
| 4 | ✓               | ✓               | ✓               | ✓               | 88.47   | 69.73  |

The result shows that \( M = 3 \) in low-layer statistics and \( M = 10 \) for top-layer features (equal to the number of classes) can reach the best results in the MNIST benchmark. Bold indicates optimal performance at the corresponding setting.
Table 8  Comparison between methods on OBJECTS benchmark using DenseNet-100

|                  | FPR95 ↓ | AUROC ↑ | AUPR ↑ |
|------------------|---------|---------|--------|
|                  | MSP     | EBO     | MDS    | ViM    | SEM    | MSP     | EBO     | MDS    | ViM    | SEM    |
| CIFAR-100        | 87.76   | 81.79   | 83.06  | 83.18  | 81.69  | 76.48   | 64.80   | 71.39  | 69.63  | 76.48  | 91.89  | 91.32  | 97.19  | 81.72  | 90.24  |
| Tiny-ImageNet    | 82.16   | 77.01   | 84.89  | 88.80  | 85.37  | 79.09   | 69.34   | 80.81  | 76.38  | 84.63  | 94.43  | 86.30  | 88.72  | 91.48  | 96.99  |
| Mean (Near-OOD)  | 84.96   | 79.40   | 83.98  | 85.99  | 83.53  | 77.79   | 67.07   | 76.10  | 73.01  | 80.56  | 93.16  | 88.81  | 92.96  | 86.60  | 93.62  |
| MNIST            | 86.61   | 84.36   | 77.04  | 87.50  | 95.50  | 71.75   | 54.83   | 75.62  | 71.81  | 78.52  | 56.36  | 39.32  | 68.32  | 67.66  | 80.05  |
| FashionMNIST     | 80.23   | 69.25   | 78.25  | 77.11  | 92.90  | 76.14   | 75.25   | 84.42  | 86.47  | 77.83  | 97.51  | 93.62  | 92.17  | 92.15  | 98.94  |
| Texture          | 87.53   | 71.19   | 70.30  | 64.83  | 78.93  | 76.45   | 74.46   | 75.41  | 78.44  | 81.47  | 93.51  | 97.44  | 90.32  | 95.25  | 99.12  |
| CIFAR-100-C      | 87.11   | 84.10   | 80.46  | 90.81  | 84.68  | 75.66   | 66.92   | 69.01  | 71.43  | 80.16  | 90.94  | 89.43  | 81.46  | 92.61  | 90.90  |
| Mean (Far-OOD)   | 85.37   | 77.23   | 76.51  | 80.06  | 88.00  | 75.00   | 67.87   | 76.12  | 77.04  | 79.50  | 84.58  | 79.95  | 83.07  | 86.92  | 92.25  |

The proposed SEM obtains a consistently better performance on most of the metrics, just as the results of its ResNet-18 counterparts in Table 3. Bold indicates optimal performance at the corresponding setting.
Table 9  Comparison between SEM and outlier exposure with negative data augmentation

| Method   | Near-OOD | Far-OOD |
|----------|----------|---------|
| OE + NDA | 32.40    | 49.14   |
| SEM      | **88.47**| **69.73**|

Bold indicates optimal performance at the corresponding setting

Table 10  SEM performance on different covariate shift levels

| Severity | Near-OOD | 1 | 2 | 3 | 4 | 5 |
|----------|----------|---|---|---|---|---|
| 0        | 83.21    | 80.70 | 75.81 | 72.56 | 65.84 | 62.12 |
| 1        | 83.45    | 74.33 | 62.35 | 52.34 | 50.11 |

We use a modified OBJECTS benchmark where the training ID is CIFAR-10 and CS-ID is CIFAR-10-C for easy control on covariate shift

5.4.6 How Covariate Shift Level affects SEM

Table 10 shows the performance of SEM scores with different covariate shift levels. We use the OBJECTS benchmark with CIFAR-10 as the training ID and CIFAR-10-C as the covariate-shifted ID. We re-generate CIFAR-10-C using different noise levels as described in Hendrycks and Dietterich (2019). Level-0 means no noise is added. The results show consistent performance degradation as the covariate shift becomes intensive. When the intensity becomes 4, SEM is generally considering the covariate shifted ID and OOD.

5.4.7 In-Distribution Classification

Table 11 shows the classification performance of MSP, MDS, and the proposed SEM score. Notice that EBO and ViM do not have special classification designs so their classification results are the same as the MSP baseline. The result shows that the SEM score has better classification performance, especially on the covariate-shifted ID datasets, which also indicates its better generalizability.

6 Discussion and Conclusion

Existing OOD detection literature has shown mostly relied on covariate shift even though they are intended to detect the semantic shift. This is very effective when test OOD data only come from the far-OOD group—where the covariate shift is large and is further exacerbated by semantic shift, so using covariate shift as a measure to detect OOD fares well. However, when it comes to near-OOD data, especially with covariate-shifted ID (i.e., data experiencing covariate shift but still belonging to the same in-distribution data), current state-of-the-art methods would suffer a significant drop in performance, as shown in the experiments.

We find the gap is caused by a shortcoming in existing evaluation benchmarks: they either exclude covariate-shifted data during testing or treat them as OOD, which is conceptually contradictory with the primary goal that a machine learning model should generalize beyond the training distribution. To fill the gap, we introduce a new problem setting that better matches the design principles of machine learning models: they should be robust in terms of good generalization to covariate-shifted datasets, and trustworthy as they also need to be capable of detecting abnormal semantic shifts.

The empirical results suggest that current state-of-the-art methods rely too heavily on covariate shift and hence could easily misclassify covariate-shifted ID data as OOD data. In contrast, our SEM score function, despite having a simple design, provides a more reliable measure for solving full-spectrum OOD detection.

In fact, in detecting samples with covariate shift, we find that a simple probabilistic model using low-level feature statistics can reach a near-perfect result.

6.1 Outlook

As the OOD detection community gets common awareness of the saturated performance problem of classic OOD benchmarks, several works have taken one step further toward the more realistic setting and proposed large-scale benchmarks (Wang et al., 2022). However, this paper shows that even under the classic MNIST/CIFAR-scale OOD benchmarks, current OOD methods in fact cannot achieve satisfactory results when the generalization ability is required. We hope that future OOD detection works could also consider the generalization capability of covariate-shifted ID data, in parallel to exploring larger-scale models and datasets.
6.2 On the Paper Title

We name the paper Full-spectrum Out-of-Distribution Detection to emphasize the main focus of our work: to consider both semantic shift and covariate shift in the OOD detection process. The term “Full-spectrum” reflects the idea that we expect the confidence value to consistently decrease from training ID, covariate-shifted ID, near-OOD, and far-OOD (ref. Fig. 1). We hope the follow-up works can further improve the OOD detection performance under the full-spectrum setting. Also, we do not use the alternative abbreviation “FS-OOD” to distinguish it from the topic of “few-shot ood detection”.

6.3 Broader Impacts

Our research aims to improve the robustness of machine learning systems in terms of the capability to safely handle abnormal data to avoid catastrophic failures. This could have positive impacts on a number of applications, ranging from consumer (e.g., AI-powered mobile phones) to transportation (e.g., autonomous driving) to medical care (e.g., abnormality detection). The new problem setting introduced in the paper includes an important but largely missing element in existing research, namely data experiencing covariate shift but belonging to the same in-distribution classes. We hope the new setting, along with the simple approach based on SEM and the findings presented in the paper, can pave the way for future research for more reliable and practical OOD detection.

Acknowledgements This study is supported by the Ministry of Education, Singapore, under its MOE AcRF Tier 2 (MOE-T2EP20221-0012), NTU NAP, and under the RIE2020 Industry Alignment Fund - Industry Collaboration Projects (IAF-ICP) Funding Initiative, as well as cash and in-kind contribution from the industry partner(s).

Data Availability Statement The datasets generated during and/or analysed during the current study are available in the OpenOOD repository, https://github.com/Jingkang50/OpenOOD.

References

Bengio, Y., Courville, A., & Vincent, P. (2013). Representation learning: A review and new perspectives. In IEEE transactions on pattern analysis and machine intelligence (TPAMI).

Bulatov, Y. (2011). NotMNIST dataset. http://yaroslavvb.blogspot.com/2011/09/notmnist-dataset.html.

Choi, H., Jang, E., & Alemi, A.A. (2018). WACIC, but why? Generative ensembles for robust anomaly detection. arXiv preprint arXiv:1810.01392.

Cimpoi, M., Maji, S., Kokkinos, I., Mohamed, S. & Vedaldi, A. (2014). Describing textures in the wild. In Proceedings of the ieee conference on computer vision and pattern recognition (CVPR).

He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep residual learning for image recognition. In Proceedings of the IEEE conference on computer vision and pattern recognition (CVPR).

Hendrycks, D. & Dietterich, T. (2019). Benchmarking neural network robustness to common corruptions and perturbations. In Proceedings of international conference on learning representations (ICLR).

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

Hendrycks, D., Mazeika, M., & Dietterich, T. (2019). Deep anomaly detection with outlier exposure. In Proceedings of international conference on learning representations (ICLR).

Huang, Gao., Liu, Zhuang., Van Der Maaten, L., & Weinberger, K.Q. (2017). Densely connected convolutional networks. In Proceedings of the IEEE conference on computer vision and pattern recognition (CVPR).

Huang, X., & Belongie, S. (2017). Arbitrary style transfer in real-time with adaptive instance normalization. In Proceedings of the IEEE conference on computer vision and pattern recognition (CVPR).

Hull, Jonathan J. (1994). A database for handwritten text recognition research. IEEE Transactions on Pattern Analysis and Machine Intelligence, 16(5), 550–554.

Krizhevsky, A., Hinton, G., et al. (2009). Learning multiple layers of features from tiny images. Citeseer.

LeCun, Y., Cortes, C. & Burges, C.J.C. (1998). The mnist database of handwritten digits. 1998. http://yann.lecun.com/exdb/mnist.

LeCun, Y. et al. (2015). Lenet-5, convolutional neural networks. http://yann.lecun.com/exdb/lenet.

Lee, K., Lee, K., Lee, H., & Shin, J. (2018). A simple unified framework for detecting out-of-distribution samples and adversarial attacks. In: Proceedings of advances in neural information processing systems (NeurIPS).

Li, Y., & Vasconcelos, N. (2020). Background data resampling for outlier-aware classification. In Proceedings of the IEEE conference on computer vision and pattern recognition (CVPR).

Liang, S., Li, Y., & Srikan, R. (2017). Enhancing the reliability of out-of-distribution image detection in neural networks. In Proceedings of international conference on learning representations (ICLR).

Lin, C., Yuan, Z., Zhao, S., Sun, P., Wang, C., & Cai, J. (2021). Domain-invariant disentangled network for generalizable object detection. In Proceedings of the IEEE international conference on computer vision (ICCV).

Liu, W., Wang, X., Owens, J., & Li Y. (2020). Energy-based out-of-distribution detection. In: Proceedings of advances in neural information processing systems (NeurIPS).

Ming, Y., Yin, H., & Li Y.(2021). On the impact of spurious correlation for out-of-distribution detection. arXiv preprint arXiv:2109.05642.

Netzer, Y., Wang, T., Coates, A., Bissacco, A., Wu, B., Ng, A.Y. (2011). Reading digits in natural images with unsupervised feature learning. In Proceedings of NIPS workshop on deep learning and unsupervised feature learning.

Peng, X., Huang, Z., Sun, X., Saenko, K. (2019). Domain agnostic learning with disentangled representations. In International conference on machine learning, pp. 5102–5112. PMLR.

Ren, J., Liu, P.J., Fertig, E., Snoek, J., Poplin, R., Depristo, M., Dillon, J., & Lakshminarayanan, B. (2019). Likelihood ratios for out-of-distribution detection. In Proceedings of advances in neural information processing systems (NeurIPS).

RSNA (2017). RSNA Pediatric Bone Age Challenge, 2017.

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

Santa Cruz, B. G., Bossa, M. N., Söller, J., & Husch, A. D. (2021). Public covid-19 x-ray datasets and their impact on model bias-a systematic review of a significant problem. Medical Image Analysis, 74, 102225.
Sastry, C.S., Oore, S. (2020). Detecting out-of-distribution examples with gram matrices. In: ICML.
Sedlmeier, A., Gabor, T., Phan, T., Belzner, L. & Claudia L.-P. (2019). Uncertainty-based out-of-distribution detection in deep reinforcement learning. arXiv preprint arXiv:1901.02219.
Serrà, J., Álvarez, D., Gómez, V., Slizovskaia, O., Núñez, J.F., Luque, J. (2020). Input complexity and out-of-distribution detection with likelihood-based generative models. In: Proceedings of international conference on learning representations (ICLR).
Sinha, A., Ayush, K., Song, K., Uzkent, B., Jin, H. & Ermon, S. (2021). Negative data augmentation. In: Proceedings of international conference on learning representations (ICLR).
Van Amersfoort, J., Smith, L., Teh, Y.W. & Gal, Y. (2020). Uncertainty estimation using a single deterministic neural network.
Vinyals, O., Ewalds, T., Bartunov, S., Georgiev, P., Vezhnevets, A.S., Yeo, M., Makhzani, A., Küttler, H., Agapiou, J., Schrittwieser, J. et al. (2017). Starcraft ii: A new challenge for reinforcement learning. arXiv preprint arXiv:2006.01174.
Vyas, A., Jamaladaka, N., Zhu, X., Das, D., Kaul, B. & Willke, T.L. (2018). Out-of-distribution detection using an ensemble of self supervised leave-out classifiers. In: Proceedings of the European conference on computer vision (ECCV).
Wang, H., Li, Z., Feng, L. & Zhang, W. (2022). Vim: Out-of-distribution with virtual-logit matching. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition.
Wang, L., Lin, Z., & Wong, A. (2020). Covid-net: A tailored deep convolutional neural network design for detection of covid-19 cases from chest x-ray images. Scientific Reports, 10(1), 1–12.
Winther, H.B., Laser, H., Gerbel, S., Maschke, S.K., Hinrichs, J.B., Vogel-Clausen, J., Wacker, F.K., Höper, M.M., & Meyer, B.C. (2020). Covid-19 image repository. URL: https://figshare.com/articles/COVID-19_Image_Repository/12275009.
Xiao, H., Rasul, K. & Vollgraf, R. (2017). Fashion-mnist: A novel image dataset for benchmarking machine learning algorithms. arXiv preprint arXiv:1708.07747.
Yang, X., He, X., Zhao, J., Zhang, Y., Zhang, S. & Xie, P. (2020). Covid-ct-dataset: A ct scan dataset about covid-19. arXiv preprint arXiv:2003.13865.
Yang, J., Wang, H., Feng, L., Yan, X., Zheng, H., Zhang, W. & Liu, Z. (2021). Semantically coherent out-of-distribution detection. In: Proceedings of the IEEE international conference on computer vision (ICCV).
Yang, J., Zhou, K., Li, Y. & Liu, Z.(2021). Generalized out-of-distribution detection: A survey. arXiv preprint arXiv:2110.11334.
Zhang, H., Cisse, M., Dauphin, Y.N., Lopez-Paz, D. (2018). Mixup: Beyond empirical risk minimization. In: Proceedings of international conference on learning representations (ICLR).
Zhou, K., Liu, Z., Qiao, Y., Xiang, T., Loy, C.C. (2022). Domain generalization: A survey. IEEE transactions on pattern analysis and machine intelligence (TPAMI).
Zhou, K., Yang, Y., Qiao, Y. & Xiang, T. (2021). Domain generalization with mixstyle. In: Proceedings of international conference on learning representations (ICLR).
Zhou, B., Lapedriza, A., Khosla, A., Oliva, A., & Torralba, A. (2017). Places: A 10 million image database for scene recognition. IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI), 40(6), 1452–1464.

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

Springer Nature or its licensor (e.g. a society or other partner) holds exclusive rights to this article under a publishing agreement with the author(s) or other rightsholder(s); author self-archiving of the accepted manuscript version of this article is solely governed by the terms of such publishing agreement and applicable law.