electronics

MDPI

Article

Enhancing Out-of-Distribution Detection Under Covariate Shifts: A Full-Spectrum Contrastive Denoising Framework

Dengye Pan, Bin Sheng and Xiaoqiang Li *

Computer Engineering and Science Department, Shanghai University, Shanghai 200444, China; pandy@shu.edu.cn (D.P.); shengbin@shu.edu.cn (B.S.)
* Correspondence: xqli@shu.edu.cn

Abstract: Out-of-distribution (OOD) detection is crucial for identifying samples that deviate from the training distribution, thereby enhancing the reliability of deep neural network models. However, existing OOD detection methods primarily address semantic shifts, where an image's inherent semantics have changed, and often overlook covariate shifts, which are prevalent in real-world scenarios. For instance, variations in image contrast, lighting, or viewpoints can alter input features while keeping the semantic content intact. To address this, we propose the Full-Spectrum Contrastive Denoising (FSCD) framework, which improves OOD detection under covariate shifts. FSCD first establishes a robust semantic boundary and then refines feature representations through fine-tuning. Specifically, FSCD employs a dual-level perturbation augmentation module to simulate covariate shifts and a feature contrastive denoising module to effectively distinguish in-distribution samples from OOD samples. Extensive experiments on three benchmarks demonstrate that FSCD achieves state-of-the-art performance, with AUROC improvements of up to 0.51% on DIGITS, 0.55% on OBJECTS, and 2.09% on COVID compared to the previous best method while also maintaining the highest classification accuracy on covariate-shifted in-distribution samples.

Keywords: out-of-distribution detection; covariate shift; full spectrum; model reliability

check for updates

Academic Editor: George A. Papakostas

Received: 2 April 2025
Revised: 28 April 2025
Accepted: 5 May 2025
Published: 6 May 2025

Citation: Pan, D.; Sheng, B.; Li, X. Enhancing Out-of-Distribution Detection Under Covariate Shifts: A Full-Spectrum Contrastive Denoising Framework. *Electronics* **2025**, *14*, 1881. https://doi.org/10.3390/electronics14091881

Copyright: © 2025 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (https://creativecommons.org/licenses/by/4.0/).

1. Introduction

Deep neural networks have been widely deployed and have achieved exceptional performance in a variety of applications, such as classification, object detection, and semantic segmentation [1–3]. However, when the models encounter samples that were not present during training, they may confidently misclassify them as known classes, leading to a significant deterioration in performance [4]. With the growing demand for intelligent detection technologies, robust models must handle out-of-distribution (OOD) inputs. This necessity has driven the development of out-of-distribution detection methods [5,6], which focus on identifying inconsistencies between training and test data.

However, the traditional setup of out-of-distribution detection still relies on the unrealistic assumption that test data only exhibit semantic shifts, overlooking the fact that in-distribution data can also undergo covariate shifts. Covariate shifts refer to changes in the statistical properties of data when the underlying semantics remain consistent. As shown in Figure 1, covariate shifts are prevalent in real-world scenarios, such as image style, noise, lighting, or viewpoints.

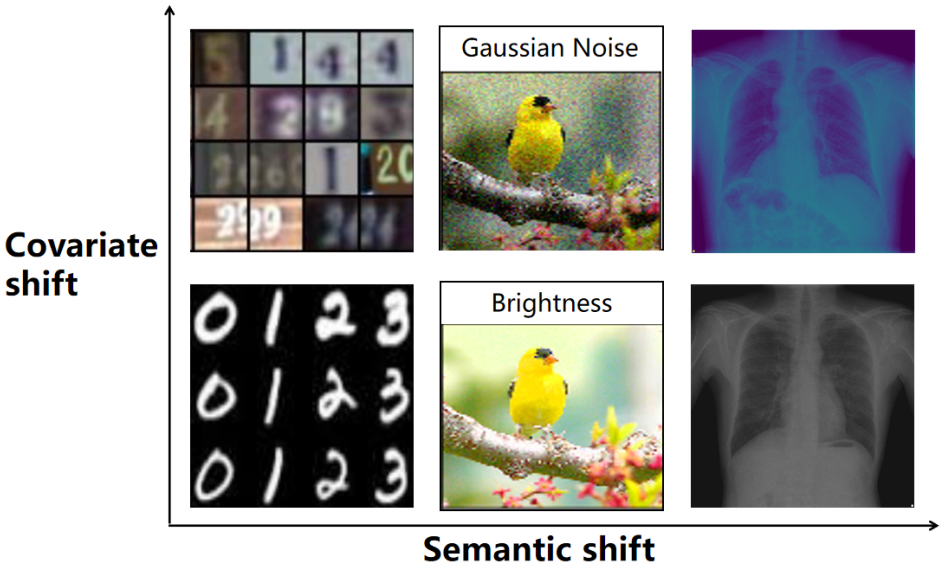


Figure 1. Illustration of covariate shifts and semantic shifts. The horizontal axis represents semantic shifts, which reflect differences across categories, while the vertical axis represents covariate shifts, which capture variations within the same category. From left to right, semantic shifts are demonstrated by changes across object categories. From top to bottom, covariate shifts are illustrated by digit images with style variations, bird images affected by noise and brightness changes, and lung X-ray images from different hospital equipment.

Zhang et al. [7] observed that out-of-distribution detection algorithms, which are designed considering only semantic shifts, suffer significant performance degradation when faced with covariate shifts. This degradation is mainly reflected in two aspects: failure to classify in-distribution (ID) samples under covariate shifts, and the tendency to misidentify them as OOD samples.

To tackle this issue, Yang et al. [8] first introduced the concept of Full-Spectrum Out-of-Distribution Detection (Fs-OOD). In the Fs-OOD task, the model encounters OOD data and ID data under covariate shifts during testing. The model is expected to classify samples exhibiting only semantic shifts as OOD while accurately identifying and classifying ID samples. Fs-OOD aims to enhance the model's robustness against various distribution changes, ensuring high performance across a wide range of practical applications.

Although several studies [8–12] have discussed covariate shifts, there is still no consensus on how to classify covariate-shifted samples. Some researchers argue that they should be considered as OOD due to their deviating distributions, while others contend that they should be excluded from OOD detection since the primary goal is to differentiate between semantic shifts. In this paper, we argue that the full-spectrum out-of-distribution detection setup is more practical, as it realistically simulates real-world conditions and enables a more comprehensive evaluation of model performance in challenging scenarios.

To address the challenge of out-of-distribution detection under covariate shifts, we propose the Full-Spectrum Contrastive Denoising (FSCD) framework. In contrast to SEM, which employs the corresponding semantic scoring function, FSCD employs a two-stage learning process—a semantic initialization stage and a fine-tuning stage—to effectively distinguish covariate-shifted ID samples from OOD samples. In the semantic initialization stage, the outlier exposure strategy is used to jointly optimize in-distribution data and auxiliary outlier samples, establishing an initial semantic decision boundary. The fine-tuning stage relies on two key components: a dual-level perturbation augmentation (DLPA) module and a feature contrastive denoising (FCD) module. DLPA introduces perturbations at both image and feature levels to enhance sample diversity and simulate covariate shifts.

Meanwhile, FCD first performs dimensionality reduction to denoise perturbed ID data and then leverages contrastive learning at the feature level to enforce semantic consistency between original and perturbed data, thereby improving the separability of ID and OOD samples in the feature space. Our proposed method (FSCD) not only improves the model's ability to handle covariate shifts but also enhances its capacity to distinguish OOD samples, marking a substantial step forward from SEM in the realm of FS-OOD detection.

The main contributions of this work are as follows:

1. We propose the Full-Spectrum Contrastive Denoising (FSCD) framework, which integrates semantic boundary initialization with fine-tuning, significantly improving OOD detection performance in scenarios with covariate shifts.
2. We propose the dual-level perturbation augmentation (DLPA) and feature contrastive denoising (FCD) modules to simulate complex covariate shifts and enable more granular differentiation of samples, thereby enhancing the model's adaptability to distributional variations.
3. The effectiveness and superiority of FSCD are validated through extensive experiments on multiple datasets, demonstrating its advantages over existing approaches in full-spectrum out-of-distribution detection and providing new insights for future research.

The structure of this paper is outlined as follows. Section 2 provides a review of existing research on out-of-distribution detection and covariate shifts, introducing recent advancements in the field and the key challenges they face. Section 3 introduces the proposed method FSCD, detailing its architecture, the dual-level perturbation augmentation module, and the feature contrastive denoising module. Section 4 presents comprehensive experimental results after evaluating FSCD on three benchmarks—DIGITS, OBJECTS, and COVID—demonstrating its state-of-the-art performance. Section 5 discusses the limitations of the approach presented in this paper and provides an outlook on possible future research directions. The conclusions are presented in Section 6.

2. Related Works

2.1. Out-of-Distribution Detection

Current methodologies in out-of-distribution detection primarily fall into two categories [13]: training-based methods and post hoc inference methods.

Training-based approaches enhance OOD awareness by explicitly integrating OOD considerations into the training process. A widely used training-based method is outlier exposure (OE) [14], which leverages the abundance of real-world unlabeled data that do not overlap with the in-distribution data. These samples form a broadly defined auxiliary outlier dataset for training. OE trains models with auxiliary OOD data by jointly optimizing with both outlier and ID samples. By exposing the model to outliers, OE encourages a more conservative inlier representation, thereby improving the detection of novel anomalies. Several studies proposed further refinements of OE-based training. OECC [15] introduces confidence calibration for outliers to improve score reliability. MixOE [16] optimizes outlier mixing ratios through gradient alignment, while BERL [17] enforces balanced regularization across different classes. Additionally, there exists a class of methods that do not require outlier data during training. For instance, SCALE [18] extends the application of scaling to the training phase, introducing an enhancement method that significantly improves OOD detection.

Post hoc inference methods boost the separation between ID and OOD by developing score functions during the inference stage. They range from the baseline maximum softmax probability (MSP) [5] to more advanced techniques. ODIN [19] enhances detection by applying temperature scaling and input perturbations to the softmax scores, while EBO [20] utilizes the energy score derived from logits to distinguish OOD samples. ReAct [21]

introduces a thresholding mechanism on feature activations to improve OOD detection, and Maximum Logit Score [22] focuses on the highest logit value for detection purposes. ATS [23] dynamically calculates a temperature value based on activations of the intermediate layers, while ExCeL [24] combines extreme and collective information within the output layer for enhanced OOD detection.

Despite notable advancements in out-of-distribution detection techniques, current methods often struggle to effectively handle scenarios involving covariate shifts. Addressing this challenge is crucial for enhancing model reliability in real-world applications.

2.2. Solutions to Covariate Shifts

A semantic shift refers to variations at the category or label level, altering the intrinsic meaning of data. In contrast, covariate shift occurs when the input distribution changes while the underlying labels remain unchanged. A typical example of a covariate shift in image processing is dehazing. Zhang et al. [25] proposed a generative adversarial and self-supervised dehazing network that uses GANs and self-supervised learning to enhance performance on natural hazy images and reduce domain shifts. Similarly, VNDHR [26] tackles nighttime haze challenges using hybrid regularization to improve illumination and reflectance. These approaches show how covariate shifts can be addressed by enhancing model generalization, thereby improving performance across various environmental conditions and image degradations.

Although many existing studies in OOD detection have discussed covariate shifts, their definitions and focus differ from those in the full-spectrum OOD detection setting. For example, ODIN [9] demonstrated that OOD samples exhibiting both semantic and covariate shifts are easier to detect than those with only one type of shift. Their work specifically examined covariate-shifted OOD samples. CSI [27] introduces augmented samples mainly for contrastive training and treats some of the augmented samples as negative examples, which is inconsistent with the definition of covariate shift in full-spectrum OOD detection. Yang et al. [8] formally introduced the full-spectrum OOD detection framework, which categorizes samples with semantic shifts as OOD, while ID samples undergoing covariate shifts are expected to be recognized as ID and correctly classified. They also proposed a semantic scoring function, SEM, comprising two probabilistic measures: one derived from high-level features that encode both semantic and non-semantic information, and another from low-level feature statistics that capture only stylistic attributes. By combining these measures, the non-semantic component is effectively eliminated, yielding a score that solely reflects semantic information. Long et al. [10] proposed an incremental shift OOD evaluation method, partitioning test samples by semantic and covariate shift intensity for refined assessment. Kyungpil [28] utilizes an adversarial mixup training method to synthesize OOD samples and evaluate OOD generalization using a distance-aware detector. Averly [11] further emphasized that the boundary between in-distribution and OOD samples is inherently model-dependent. In this paper, we follow the same experimental setup as SEM, while proposing a more robust framework that improves upon their approach through enhanced modeling and algorithmic design.

Another line of work argues that test datasets should exclude covariate-shifted samples to maintain the fundamental goal of OOD detection. A clean ImageNet-OOD dataset [12] was created to support this perspective, which only has semantic shifts to ensure a more controlled evaluation environment.

However, the setup of Fs-OOD holds greater practical significance, particularly in applications such as autonomous driving, where environmental variations, such as illumination changes, seasonal vegetation differences, and weather conditions, must be accounted for to ensure reliable decision-making. This paper adopts the same setting.

3. Proposed Method

3.1. Overall Architecture

The proposed FSCD framework is based on the concept of outlier exposure and consists of two stages: semantic boundary initialization and fine-tuning.

In the semantic boundary initialization stage, also referred to as the pretraining stage, the model is trained on both ID samples and auxiliary outlier data to establish an initial classification boundary and capture the semantic features of each class. The primary objective during this stage is to learn fundamental classification abilities.

The fine-tuning stage is specifically designed for the Fs-OOD task, as shown in Figure 2. In this phase, the input consists of in-distribution samples, auxiliary outliers, and augmented ID samples generated via both image-level and feature-level perturbations. These samples are passed through a shared feature extractor to obtain the corresponding feature representations: $\boldsymbol{F}_{\text{ID}}$ for clean ID data, $\boldsymbol{F}_{\text{Aug}}$ for image-perturbed samples, $\boldsymbol{F}_{\text{Aug}'}$ for feature-perturbed samples, and $\boldsymbol{F}_{\text{OOD}}$ for outliers.

Since perturbations may introduce noise into ID representations, FSCD first applies feature denoising to suppress redundant low-level variations. Then, contrastive learning is employed to align $\boldsymbol{F}_{\text{ID}}$, $\boldsymbol{F}_{\text{Aug}}$, and $\boldsymbol{F}'_{\text{Aug}}$ while enforcing separation from $\boldsymbol{F}_{\text{OOD}}$. This combination of denoising and contrastive learning encourages semantic consistency across perturbed variants and improves the model's ability to discriminate between in- and out-of-distribution samples under covariate shifts.

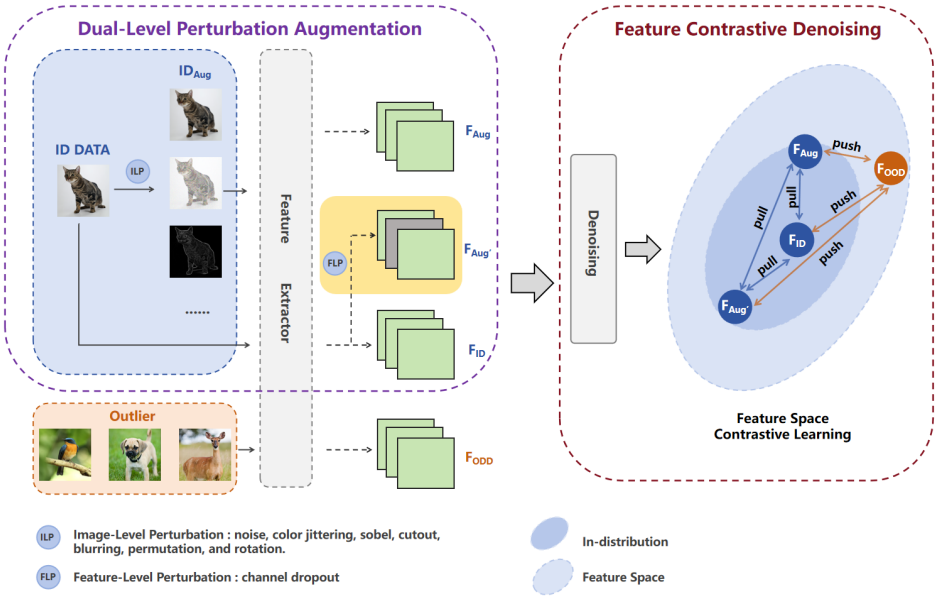


Figure 2. Illustration of the proposed fine-tuning stage in the FSCD framework. The left part shows the dual-level perturbation augmentation module, where ID data undergo both image-level and feature-level perturbations, resulting in two perturbed features: $\mathbf{F}_{\text{Aug}}$ and $\mathbf{F}_{\text{Aug}'}$. Outlier samples are processed through the feature extractor to produce $\mathbf{F}_{\text{OOD}}$. The DLPA module thus outputs clean ID features $\mathbf{F}_{\text{ID}}$, image-perturbed features $\mathbf{F}_{\text{Aug}}$, feature-perturbed features $\mathbf{F}_{\text{Aug}'}$, and outlier features $\mathbf{F}_{\text{OOD}}$, which serve as the input to the feature contrastive denoising module. The FCD module performs denoising and contrastive learning in the feature space to pull $\mathbf{F}_{\text{ID}}$, $\mathbf{F}_{\text{Aug}}$, and $\mathbf{F}_{\text{Aug}'}$ closer while pushing them away from $\mathbf{F}_{\text{OOD}}$. These pull-and-push operations aim to minimize intra-class variation and maximize inter-distribution separation, enhancing the model's robustness against covariate shifts.

3.2. Semantic Boundary Initialization Stage

During the semantic boundary initialization stage, FSCD leverages both in-distribution samples and auxiliary outliers to learn a robust semantic classification boundary. This stage

corresponds to a standard classifier pretraining process, in which the model is trained to accurately predict class labels for ID samples while simultaneously learning to reject outliers by assigning them to a designated rejection class. Although conceptually simple, this stage is essential for providing a strong initialization that facilitates subsequent perturbation modeling and contrastive denoising efforts. This phase enables the model to acquire a fundamental classification capability and form a coarse semantic separation between ID and OOD data in the feature space, laying the foundation for the fine-tuning stage.

The optimization objective in this stage is defined as follows. A rejection class [29,30] for outliers is introduced as the $(N+1)$-th class, where N denotes the number of ID categories and $\boldsymbol{x'}$ represents the outlier samples drawn from the auxiliary outlier training set $\mathcal{D}_{\text{out}}^{\text{train}}$. ℓ denotes the cross-entropy loss for classification, $f(x)$ is the predicted probability distribution for input x, and y is the ground-truth label for the ID samples. The set $\mathcal{D}_{\text{in}}^{\text{train}}$ denotes the in-distribution training set, and the hyperparameter λ balances the loss components, ensuring a trade-off between accurate classification of ID data and reliable detection of outliers.

$$\mathcal{L}_{\text{SBI}} = \mathbb{E}_{(x,y)\sim\mathcal{D}_{\text{in}}^{\text{train}}}[\ell(f(x),y)] + \lambda \cdot \mathbb{E}_{x'\sim\mathcal{D}_{\text{out}}^{\text{train}}}[\ell(f(\boldsymbol{x'}),N+1)] \tag{1}$$

3.3. Fine-Tuning Stage

The fine-tuning stage consists of two key modules: dual-level perturbation augmentation (DLPA) and feature contrastive denoising (FCD), designed to enhance the model's ability to distinguish covariate-shifted ID samples from OOD samples. The following sections detail the composition, functionality, and optimization objectives of each module.

3.3.1. Dual-Level Perturbation Augmentation Module

To address the misclassification of in-distribution samples that are affected by covariate shifts, we propose DLPA.

In the full-spectrum OOD detection setting, models are expected to handle a wide range of unknown distribution shifts during testing, which makes it impractical to rely on a specific optimization objective for generating samples. Instead, DLPA applies a variety of perturbations directly to training samples, mimicking real-world covariate shifts in a non-parametric, augmentation-based manner. As illustrated in Figure 2, DLPA introduces perturbations to in-distribution samples at both the image and feature levels [27,31,32]. The primary objective is to generate a diverse set of perturbed samples, which enhances the model's ability to generalize across various covariate shifts. By exposing the model to a broad range of perturbations, DLPA facilitates a more comprehensive understanding of data variations, improving its robustness against complex distribution shifts.

At the image level, DLPA builds upon the setup of CSI [27] and applies a series of data augmentation techniques, including noise, color jittering, cutout, Sobel, blurring, permutation, and rotation. While CSI argues that certain augmentations may distort the original data distribution, we contend that they effectively preserve the core semantic information of in-distribution samples while simulating diverse covariate shifts. For instance, Gaussian noise is introduced to model sensor noise and environmental interference, while color jittering randomly adjusts hue, saturation, and brightness to account for variations in lighting conditions and imaging devices. Brightness adjustment alters overall luminance, simulating changes in ambient lighting. By incorporating these perturbations, DLPA significantly enhances training diversity, improving the model's robustness to real-world covariate shifts.

After the image-level perturbation, the samples are processed through the feature extractor and subjected to feature-level perturbation. It should be noted that simple

channel dropout performs sufficiently well [31,32], so we apply it to feature representations to enhance robustness. This regularization strategy encourages the model to learn more discriminative and invariant features, effectively widening its decision boundaries in the feature space.

DLPA retains a clean branch for PCA-based dimensionality reduction and denoising training, serving as a supervision signal for learning stable representations. It applies perturbations at different levels separately on the clean ID branch, rather than applying feature perturbations to already perturbed data, ensuring independent operation on the clean branch. This design allows the model to accurately capture the effects of perturbations, thereby enhancing robustness and generalization. After processing through DLPA, the final outputs consist of the original in-distribution features $\boldsymbol{F}_{\mathrm{ID}}$; the feature-perturbed representations $\boldsymbol{F}_{\mathrm{Aug}'}$, which capture transformations in the feature space; and the image-perturbed representations $\boldsymbol{F}_{\mathrm{Aug}}$, which reflect variations in visual appearance. This structured perturbation strategy strengthens generalization on in-distribution samples, ensuring stable performance under complex shifts.

3.3.2. Feature Contrastive Denoising Module

FCD is designed to enhance the model's robustness against ID perturbations while improving its capacity to distinguish OOD samples. By leveraging both feature denoising and contrastive learning, FCD facilitates a more structured and discriminative feature representation.

During the denoising phase, a dimensionality reduction technique is applied to filter out perturbational noise while preserving essential semantic information. While high-dimensional features extracted from deep networks contain rich class-discriminative cues, they also capture irrelevant variations and noise, particularly in perturbed ID samples. To address this, we use Principal Component Analysis (PCA) to project features onto a lower-dimensional subspace that retains high-variance components while filtering out low-variance directions associated with noise. Specifically, we retain 99% of the total variance to ensure that only the most informative components are preserved. To avoid interference from OOD samples, the transformation is learned exclusively from clean ID features, ensuring a more reliable representation of in-distribution variation.

To refine feature representations, we incorporate contrastive learning into Fs-OOD. By constructing positive and negative sample pairs, this method promotes feature consistency for ID samples while maximizing the separation between ID and OOD distributions. We define three optimization objectives to govern the relationships among ID, perturbed ID, and OOD samples. First, the model minimizes the feature distance between an ID sample and its perturbed counterpart to learn invariant features. Second, it maximizes the feature distance between the perturbed ID and OOD samples, as well as between the ID and OOD samples, ensuring clear distributional boundaries. This structured optimization improves the model's generalization to unseen data distributions, enhancing detection accuracy and robustness in dynamic environments.

The loss function $\mathcal{L}_{\mathrm{CD}}$ builds upon the conventional L2 loss [33] by introducing a margin constraint for negative sample pairs. This loss encourages alignment between the ID samples and their perturbed variants (positive pairs) while enforcing a minimum margin on negative pairs (ID/OOD and perturbed/OOD) when their feature distances fall below a threshold. The mathematical formulation is given by

$$\begin{aligned}\mathcal{L}_{\mathrm{CD}} = \mathbb{E}_{\boldsymbol{x}\sim\mathcal{D}_{\mathrm{in}}^{\mathrm{train}}}\left[\|f(\boldsymbol{x}) - f\left(G_{\mathrm{aug}}(\boldsymbol{x})\right)\|_2^2\right] \\ - k\cdot\left(\mathbb{E}_{\substack{\boldsymbol{x}\sim\mathcal{D}_{\mathrm{in}}^{\mathrm{train}}\\ \boldsymbol{x}'\sim\mathcal{D}_{\mathrm{out}}^{\mathrm{train}}}}\left[\max\left(0, m - \|f(\boldsymbol{x}) - f(\boldsymbol{x}')\|_2^2\right)\right]\right. \\ \left. + \mathbb{E}_{\substack{\boldsymbol{x}\sim\mathcal{D}_{\mathrm{in}}^{\mathrm{train}}\\ \boldsymbol{x}'\sim\mathcal{D}_{\mathrm{out}}^{\mathrm{train}}}}\left[\max\left(0, m - \|f\left(G_{\mathrm{aug}}(\boldsymbol{x})\right) - f(\boldsymbol{x}')\|_2^2\right)\right]\right).\end{aligned} \tag{2}$$

Here, G_{aug} denotes the augmentation function applied to the input data, and m represents the margin threshold, which is a hyperparameter that determines the minimum distance that features of negative sample pairs should maintain. The parameter k is a scaling factor that controls the relative importance of the second term.

Since one of the primary objectives in OOD detection is to maintain classification accuracy on in-distribution samples, the fine-tuning phase continues to incorporate classification learning for ID data. The overall loss function $\mathcal{L}_{\mathrm{FT}}$ for the fine-tuning phase is formulated as

$$\mathcal{L}_{\mathrm{FT}} = \mathbb{E}_{(x,y)\sim\mathcal{D}_{\mathrm{in}}^{\mathrm{train}}}[\ell(f(x), y)] + \lambda_1\cdot\mathcal{L}_{\mathrm{CD}}. \tag{3}$$

The parameter λ_1 is a scaling factor that controls the relative weight between the classification loss and the contrastive denoising loss during the fine-tuning phase. The total optimization objective, while preserving ID classification accuracy, enforces compact feature representations for ID samples and ensures sufficient separation from OOD samples, thereby enhancing the model's discriminative power and robustness to distribution shifts.

4. Experiments

In this section, we first introduce the implementation of the experiments, providing a detailed description of the datasets, experimental settings, and evaluation metrics. Subsequently, we report the experimental results of the proposed algorithm on several standard benchmark datasets. To further validate the effectiveness of the proposed modules, systematic ablation studies are conducted, alongside an in-depth analysis of key hyperparameters.

4.1. Datasets

Following the settings in SEM [8], this section conducts experiments on three full-spectrum OOD detection benchmarks: DIGITS, OBJECTS, and COVID. Unlike conventional OOD evaluations, the ID test set includes covariate shift datasets, which share the same class labels and semantic features as the ID data but differ in collection methods, image styles, and statistical characteristics.

DIGITS: The ID dataset of DIGITS is MNIST [34], a grayscale handwritten digit dataset. The ID test set comprises the original MNIST test set along with covariate-shifted samples from SVHN and USPS [35]. SVHN consists of digit images captured from house number plates in real-world scenes, introducing variations in background, illumination, and orientation. The USPS dataset, collected from postal mail, features diverse handwriting styles and smaller image dimensions, which pose additional challenges for recognition. The OOD test set includes notMNIST and FashionMNIST [36], which share visual similarities with MNIST. The OOD test set also includes Texture [37], CIFAR-10 [38], Tiny-ImageNet [39], and Places365 [40], which exhibit distinct semantic and structural differences.

OBJECTS: The ID dataset is CIFAR-10 [38], which contains color images from 10 object categories. The ID test set includes the standard CIFAR-10 test set along with covariate-shifted samples from CIFAR-10-C and ImageNet-10 [41]. CIFAR-10-C introduces various corruption types (e.g., noise, blur, and contrast changes) at different severity levels, simulating real-world degradation. ImageNet-10 consists of fine-grained categories from

ImageNet-22K that semantically correspond to CIFAR-10 classes. The OOD test set includes CIFAR-100, Tiny-ImageNet, MNIST, FashionMNIST, Texture, and CIFAR-100-C.

COVID: This benchmark evaluates OOD detection in medical imaging. The ID dataset is BIMCV [42], a lung X-ray dataset. The covariate shift test set is introduced using ActMed [43] and Hannover [44], which contain X-ray images collected from different hospitals. The OOD test set includes CT-SCAN [45], XRayBone [46], MNIST, CIFAR-10, Texture, and Tiny-ImageNet.

Auxiliary outlier training set: TinyImages80M [47] is utilized as the auxiliary outlier training set, designated as $\mathcal{D}_{\text{out}}^{\text{train}}$. It is an image classification dataset provided by Stanford University. As a reduced version of ImageNet, it has become a popular benchmark due to its relatively small scale and diverse class distribution. Since CIFAR-10 and CIFAR-100 are labeled subsets of the TinyImages80M dataset, we follow the same deduplication procedure as in [9] and remove images from the dataset that belong to CIFAR-10 or CIFAR-100.

4.2. Experimental Settings

For the network architecture, appropriate structures were chosen for different benchmarks. Specifically, LeNet-5 was utilized for the DIGITS benchmark, while ResNet-18 was adopted for both the OBJECTS and COVID benchmarks. All models were trained using the stochastic gradient descent (SGD) optimizer with a momentum of 0.9 and a weight decay of 5×10^{-4}. For the DIGITS and OBJECTS benchmarks, the initial learning rate was set to 0.1 and was gradually decayed using a cosine annealing schedule over a total of 100 training epochs. For the COVID benchmark, the initial learning rate was set to 0.001, and the total number of training epochs was increased to 200 to accommodate the higher complexity of this benchmark. Additionally, during the fine-tuning phase, the learning rate was adjusted to 0.005, and the total number of fine-tuning epochs was set to 10. A fixed batch size of 128 was used across all benchmarks to balance computational efficiency and performance.

For hyperparameter configuration, we followed the commonly used values in outlier exposure methods. Specifically, λ was set to 0.5, λ_1 was set to 0.3, and k was set to 0.5. For the parameter m, a dynamic calculation strategy was employed: within each batch, the feature distances for different sample pairs were computed, and the 95th percentile was selected as the threshold to adapt to changes in the feature distribution. In particular, two types of feature distances were computed: one between the perturbed samples and the OOD samples, and another between the ID samples and the OOD samples; the maximum of these two distances was then taken as the final threshold.

In the experiments, we compared FSCD with several representative OOD detection methods, including MSP [5], EBO [20], MDS [48], ViM [49], IPL [50], and the current state-of-the-art approach, SEM [8]. MSP is a widely adopted baseline that detects OOD samples based on the maximum softmax probability. Despite its simplicity, it remains a strong and effective reference method due to its practicality and broad applicability. EBO distinguishes OOD samples using energy scores computed from the model's logits. Incorporating temperature scaling and employing negative energy as the detection metric alleviates the overconfidence issue commonly observed in deep neural networks. MDS introduces a generative classifier under the linear discriminant analysis (LDA) assumption. It models class-conditional Gaussian distributions over both lower- and higher-level features of deep networks and defines a confidence score based on the Mahalanobis distance, effectively capturing feature-space deviations of OOD inputs. ViM uses a novel scoring strategy termed Virtual-logit Matching, which combines a class-agnostic component from the feature space with class-dependent logits. This hybrid formulation enhances OOD detection by jointly modeling semantic similarity and uncertainty. IPL is a novel method that identifies and concentrates on intrinsic parameters that capture semantic correlations, while

reducing the negative impacts of covariance overfitting, to boost the model's robustness and generalization in OOD detection tasks. SEM, the most recent and relevant method, formally introduces the full-spectrum OOD detection task. In our experiments, we conducted comprehensive comparisons with SEM to demonstrate the advantages and robustness of FSCD under various covariate and semantic shift scenarios.

4.3. Evaluation Metrics

To rigorously evaluate the efficacy of full-spectrum out-of-distribution detection, we employed four widely recognized metrics: the Area Under the Receiver Operating Characteristic Curve (AUROC), the Area Under the Precision–Recall Curve (AUPR), the false positive rate at a 95% true positive rate (FPR95), and accuracy (ACC) on the ID test set, including covariate-shifted samples. AUROC measures the model's ability to distinguish between classes by plotting the true positive rate against the false positive rate at various threshold settings, while AUPR focuses on the trade-off between precision and recall, providing a more informative view of performance for imbalanced datasets. FPR95 quantifies the false positive rate when the true positive rate reaches 95%, offering insight into model specificity. ACC evaluates the model's accuracy on the ID test set, reflecting its generalization ability under distribution shifts. A robust model is expected to achieve high AUROC and AUPR values and a low FPR95 value and maintain an ACC comparable to standard ID performance.

4.4. Results and Analysis

Overall, our proposed FSCD effectively distinguished various OOD datasets, achieving excellent performance in both Near-OOD tasks, where the OOD samples were similar to the ID samples, and Far-OOD tasks, where the style and content differed significantly from the ID samples. In particular, FSCD outperformed existing approaches on the AUROC and AUPR metrics across all benchmarks. Furthermore, compared to SEM, FSCD exhibited a lower FPR95, indicating enhanced robustness in OOD detection.

Table 1 compares the proposed FSCD method with existing methods for full-spectrum OOD detection using the AUROC metric. FSCD achieved state-of-the-art performance across all three benchmarks, attaining mean AUROC scores of 76.49%, 78.07%, and 90.11% on the DIGITS, OBJECTS, and COVID datasets, respectively, consistently outperforming previous approaches. Notably, FSCD demonstrated a balanced superiority in both Near-OOD and Far-OOD scenarios. It improved the AUROC by 0.73–1.45% over the second-best method (SEM) on DIGITS and OBJECTS while maintaining competitive performance in Near-OOD scenarios, with AUROC values of 88.54% and 76.78%. Particularly compelling were the results on the COVID benchmark, where FSCD achieved a substantial 3.22% average improvement over SEM. These systematic improvements validate FSCD's effectiveness in addressing diverse distribution shifts.

Table 2 evaluates FSCD's performance on the AUPR metric for full-spectrum OOD detection. FSCD achieved the highest average AUPR scores of 92.34%, 90.15%, and 83.66% on the DIGITS, OBJECTS, and COVID benchmarks, respectively, demonstrating a superior precision–recall balance compared to existing approaches. Compared to SEM, FSCD improved the Near-OOD performance by 0.51% on DIGITS and 0.68% on OBJECTS. The results on COVID were compelling as well: FSCD attained near-perfect detection rates of 99.80% on CT-SCAN and 99.46% on XRayBone and maintained robust performance in Far-OOD scenarios. These results emphasize FSCD's dual capability to precisely distinguish Near-OOD samples while preserving reliability across diverse shifts.

Table 1. Comparison of FSCD with existing methods on the AUROC metric across three benchmarks. The best results are highlighted in bold, and the second-best results are underlined.

Dataset	MSP	EBO	MDS	ViM	SEM	IPL	FSCD
-DIGITS (Training ID: MNIST; Covariate-shifted ID: USPS and SVHN)							
notMNIST	32.54	25.49	79.10	85.91	**96.74**	95.32	96.28
FashionMNIST	39.71	37.64	60.42	66.45	80.20	79.28	**80.79**
Mean (Near-OOD)	36.12	31.56	69.76	76.18	88.47	87.29	**88.54**
Texture	64.34	65.02	72.42	70.32	74.45	73.97	**74.73**
CIFAR-10	52.22	50.95	67.96	63.23	69.29	70.33	**71.09**
Tiny-ImageNet	52.94	51.89	64.31	62.34	67.54	67.64	**68.10**
Places365	50.22	48.95	65.42	65.32	67.63	65.49	**67.91**
Mean (Far-OOD)	54.93	54.20	67.53	65.30	69.73	69.36	**70.46**
Mean (Average)	48.66	46.66	68.27	68.93	75.98	75.34	**76.49** $\pm$ 0.35
-OBJECTS (Training ID: CIFAR-10; Covariate-shifted ID: CIFAR-10-C and ImageNet-10)							
CIFAR-100	70.17	63.85	72.05	70.12	74.70	73.19	**75.91**
Tiny-ImageNet	72.92	67.97	72.94	73.24	76.76	76.31	**77.65**
Mean (Near-OOD)	71.55	65.91	72.50	71.68	75.73	74.75	**76.78**
MNIST	66.98	54.55	77.04	73.45	**75.69**	73.13	74.82
FashionMNIST	73.78	76.50	80.33	**81.01**	79.40	80.04	80.33
Texture	74.18	68.63	72.02	73.40	79.69	77.41	**79.72**
CIFAR-100-C	74.12	68.37	68.13	73.32	78.89	79.86	**80.01**
Mean (Far-OOD)	72.27	67.01	74.38	75.30	78.42	77.61	**78.72**
Mean (Average)	72.03	66.65	73.75	74.09	77.52	76.66	**78.07** $\pm$ 0.27
-COVID (Training ID: BIMCV; Covariate-shifted ID: ActMed and Hannover)							
CT-SCAN	11.31	13.14	81.21	89.23	**99.51**	95.69	97.33
XRayBone	32.08	77.80	78.72	90.37	94.97	95.00	**96.12**
Mean (Near-OOD)	21.70	45.47	79.96	89.80	**97.24**	95.35	96.73
MNIST	24.89	99.91	80.81	98.32	**100.00**	98.75	99.80
CIFAR-10	41.12	45.23	**77.05**	46.30	52.50	68.93	72.40
Texture	22.63	34.95	89.84	82.80	90.94	87.43	**91.24**
Tiny-ImageNet	30.26	32.69	81.99	79.32	83.42	82.29	**83.75**
Mean (Far-OOD)	29.73	53.20	82.42	76.69	81.72	84.35	**86.80**
Mean (Average)	27.05	50.62	81.60	81.06	86.89	88.02	**90.11** $\pm$ 0.81

The COVID benchmark analysis revealed a critical advantage of FSCD in practical deployment scenarios. While MDS achieved notable AUPR values on certain Far-OOD datasets, its mean FPR95 of 90.98% significantly exceeded FSCD's 26.77%, as evidenced in Table 3. This substantial 64.21% reduction in the false positive rate demonstrates FSCD's capability to maintain detection sensitivity while effectively suppressing erroneous alerts, which is particularly crucial for medical imaging applications. The consistent superiority across benchmarks was further validated by FSCD's lowest average FPR95 scores on both DIGITS and COVID, demonstrating its enhanced robustness against data perturbations in real-world environments.

As shown in Table 4, FSCD demonstrated robust classification performance on both the ID test set and the covariate-shifted ID test set. Notably, this study follows the experimental settings of SEM and reports classification accuracy only on the DIGITS and OBJECTS benchmarks. Because the datasets in the COVID benchmark are divided into positive and negative classes, with all covariate-shifted test samples belonging to the positive class, the accuracy metric is limited in its ability to fully capture real-world performance. Therefore, accuracy results are reported only for the DIGITS and OBJECTS benchmarks.

Table 2. Comparison of FSCD with existing methods on the AUPR metric across three benchmarks. The best results are highlighted in bold, and the second-best results are underlined.

Dataset	MSP	EBO	MDS	ViM	SEM	IPL	FSCD
-DIGITS (Training ID: MNIST; Covariate-shifted ID: USPS and SVHN)							
notMNIST	67.33	63.97	90.60	92.23	**98.54**	97.28	<u>97.72</u>
FashionMNIST	82.40	81.57	88.84	89.05	94.38	<u>95.73</u>	**96.23**
Mean (Near-OOD)	74.87	72.77	89.72	90.64	96.46	<u>96.51</u>	**96.97**
Texture	94.40	94.47	95.81	96.01	<u>96.12</u>	95.26	**96.78**
CIFAR-10	87.26	86.36	91.74	90.52	<u>92.06</u>	91.75	**93.10**
Tiny-ImageNet	87.51	86.72	90.71	90.13	<u>91.58</u>	91.03	**92.09**
Places365	67.11	65.41	76.64	73.12	77.61	<u>77.64</u>	**78.12**
Mean (Far-OOD)	84.07	83.24	88.73	87.45	<u>89.34</u>	88.92	**90.02**
Mean (Average)	81.00	79.75	89.06	88.51	<u>91.72</u>	91.45	**92.34** $\pm$ 0.33
-OBJECTS (Training ID: CIFAR-10; Covariate-shifted ID: CIFAR-10-C and ImageNet-10)							
CIFAR-100	88.28	83.51	89.42	82.93	<u>90.64</u>	89.66	**91.86**
Tiny-ImageNet	90.04	86.30	89.96	83.57	<u>91.86</u>	90.97	**92.00**
Mean (Near-OOD)	89.16	84.91	89.69	83.25	<u>91.25</u>	90.31	**91.93**
MNIST	52.66	34.14	65.31	60.33	**76.61**	74.07	<u>74.58</u>
FashionMNIST	90.15	89.80	92.28	90.47	**93.14**	92.25	<u>93.02</u>
Texture	93.34	89.51	88.46	90.55	<u>95.48</u>	95.11	**96.30**
CIFAR-100-C	89.74	85.54	82.97	90.27	<u>92.07</u>	91.73	**93.14**
Mean (Far-OOD)	81.47	74.75	82.25	82.91	**89.33**	88.29	<u>89.26</u>
Mean (Average)	84.04	78.13	84.73	83.02	<u>89.97</u>	88.97	**90.15** $\pm$ 0.12
-COVID (Training ID: BIMCV; Covariate-shifted ID: ActMed and Hannover)							
CT-SCAN	52.92	53.34	94.31	88.90	**99.80**	<u>97.84</u>	**99.80**
XRayBone	76.95	91.68	96.67	97.84	<u>98.95</u>	98.70	**99.46**
Mean (Near-OOD)	64.94	72.51	95.49	93.37	<u>99.37</u>	98.27	**99.63**
MNIST	1.07	95.90	81.11	97.27	**100.00**	98.26	<u>99.70</u>
CIFAR-10	8.73	9.77	**61.14**	18.74	11.27	38.21	<u>46.03</u>
Texture	11.43	13.14	<u>85.50</u>	57.21	64.71	83.01	**86.66**
Tiny-ImageNet	7.42	7.65	**77.94**	45.31	31.19	69.38	<u>70.31</u>
Mean (Far-OOD)	7.16	31.62	**76.42**	54.63	51.79	48.14	<u>75.68</u>
Mean (Average)	26.42	45.25	<u>82.78</u>	67.55	67.65	64.85	**83.66** $\pm$ 0.47

On datasets with covariate shifts, the overall accuracy of FSCD was significantly higher than that of the baseline methods, thereby validating the effectiveness and robustness of our approach in handling covariate shifts. More specifically, on the original MNIST and CIFAR-10 test sets, FSCD achieved classification accuracies of 99.8% and 95.3%, respectively, indicating that the integration of denoising and contrastive learning mechanisms does not compromise the model's ability to recognize the original ID data.

Although FSCD performed slightly worse than SEM on the SVHN test set, it remained competitive overall. On the more challenging CIFAR-10-C test set, which exhibits severe covariate shifts, FSCD achieved an accuracy of 63.2%, surpassing SEM and demonstrating enhanced robustness in the presence of complex covariate shifts.

Additionally, we report the deviations of FSCD across different benchmarks, showing that FSCD exhibited consistent performance with minimal fluctuations across experimental runs.

Table 3. Comparison of FSCD with existing methods on the FPR95 metric across three benchmarks. The best results are highlighted in bold, and the second-best results are underlined.

Dataset	MSP	EBO	MDS	ViM	SEM	IPL	FSCD
-DIGITS (Training ID: MNIST; Covariate-shifted ID: USPS and SVHN)							
notMNIST	99.97	99.99	78.83	66.34	**10.93**	15.68	11.01
FashionMNIST	99.90	99.98	94.68	93.78	68.63	67.10	**65.30**
Mean (Near-OOD)	99.93	99.98	86.75	80.06	39.78	41.39	**38.16**
Texture	94.89	98.40	**87.46**	92.34	90.90	90.02	89.72
CIFAR-10	98.01	99.62	95.47	93.60	91.57	90.39	**90.04**
Tiny-ImageNet	97.98	99.58	96.20	96.18	93.39	93.30	**92.17**
Places365	98.68	99.65	98.06	98.02	94.15	95.74	**92.33**
Mean (Far-OOD)	97.39	99.31	94.30	95.04	92.50	92.36	**91.07**
Mean (Average)	98.24	99.54	91.78	90.04	74.93	75.37	**73.43** $\pm$ 0.64
-OBJECTS (Training ID: CIFAR-10; Covariate-shifted ID: CIFAR-10-C and ImageNet-10)							
CIFAR-100	89.44	83.84	86.28	**82.44**	86.96	84.31	84.24
Tiny-ImageNet	88.22	**81.58**	87.45	87.32	86.59	96.17	85.55
Mean (Near-OOD)	88.83	**82.71**	86.87	84.89	86.77	90.24	84.89
MNIST	93.54	92.23	**84.59**	90.12	99.70	95.39	94.81
FashionMNIST	88.08	**72.40**	77.17	75.32	93.72	78.62	76.51
Texture	85.64	75.57	72.98	**70.92**	82.15	79.38	72.60
CIFAR-100-C	87.26	83.64	85.53	89.32	83.92	83.02	**82.17**
Mean (Far-OOD)	88.63	80.96	**80.07**	81.42	89.87	84.10	81.52
Mean (Average)	88.70	**81.54**	83.33	82.57	88.84	86.15	82.65 $\pm$ 0.57
-COVID (Training ID: BIMCV; Covariate-shifted ID: ActMed and Hannover)							
CT-SCAN	99.80	97.35	99.39	12.56	**2.24**	9.06	8.72
XRayBone	97.00	42.00	100.00	23.35	14.50	14.17	**10.07**
Mean (Near-OOD)	98.40	69.67	99.69	17.96	**8.37**	11.62	9.39
MNIST	98.30	0.35	100.00	1.69	**0.00**	1.24	0.78
CIFAR-10	96.32	94.67	98.02	89.61	85.58	85.94	**82.09**
Texture	98.39	87.06	56.38	40.85	27.57	28.40	**13.28**
Tiny-ImageNet	97.78	92.73	92.11	73.96	**44.99**	45.92	45.70
Mean (Far-OOD)	97.70	68.70	86.63	51.38	39.54	40.38	**35.46**
Mean (Average)	97.93	69.03	90.98	40.34	29.15	30.78	**26.77** $\pm$ 0.98

Table 4. Comparison of existing methods' classification accuracies on the ID datasets (MNIST and CIFAR-10) and their corresponding covariate shift datasets. The best results are highlighted in bold, and the second-best results are underlined.

Dataset	MNIST	USPS	SVHN	CIFAR-10	ImageNet-10	CIFAR-10-C
MSP	99.6	60.2	22.1	94.3	80.0	52.3
MDS	99.0	61.4	29.1	95.1	82.2	50.4
SEM	99.4	75.2	**40.2**	94.2	**85.7**	61.5
Ours	**99.8**	**76.0**	40.0	**95.3**	**85.7**	**63.2**

4.5. Effectiveness of the Modules

To evaluate the influence of each component within FSCD, detailed ablation experiments were conducted on the DIGITS benchmark, as summarized in Table 5. The table reports the average performance of various module combinations across six test sets, highlighting metrics such as AUROC, AUPR, FPR95, and ACC.

The first row displays the performance when only the pretraining phase was employed. The results indicate that the model performed poorly in both OOD detection and classification tasks under covariate shifts. This underscores that pretraining alone merely establishes initial semantic boundaries and is insufficient for addressing Fs-OOD tasks.

The second row shows the model's performance when it was fine-tuned from scratch, with all modules in the fine-tuning phase applied from the beginning. Although the

AUROC and FPR95 metrics were relatively prominent, the model failed to learn initial boundary information, resulting in a sharp decline in classification accuracy (with AUPR approaching 50%). This demonstrates the complementary nature and sequential necessity of pretraining and fine-tuning. Based on these observations, subsequent ablation experiments were conducted following the pretraining phase.

Notably, within the proposed FSCD framework, DLPA and FCD form a closely integrated process, with the output of FSCD directly serving as the input for FCD. In this context, the analysis focused on key aspects, such as perturbations at different levels, feature contrast, and reduction denoising, to explore their individual impact on the final results.

The results in the third and fourth rows indicate that employing perturbations at either the image or feature level alone leads to performance improvements. Since image-level perturbations are more diverse, while feature-level perturbations rely solely on dropout techniques, these findings align with expectations.

Next, the roles of contrastive learning and dimensionality reduction denoising within the FCD module were examined. In FSCD, dimensionality reduction denoising serves as a preliminary step, while the primary performance optimization is driven by subsequent contrastive differentiation. To conduct a comprehensive ablation study, these two components were evaluated separately.

The fifth row presents the results when only denoising was applied to perturbed images. Since dimensionality reduction primarily filters out shallow and redundant noise, and the perturbed images generated through DLPA are not further refined by FCD, the performance improvement over pretraining remained marginal. The sixth row shows the outcomes of directly applying contrastive push-pull between perturbed data, ID data, and outliers. By reducing the feature distance between ID samples and their corresponding perturbed versions while increasing the separation between ID samples and auxiliary outliers, the model's capacity to distinguish out-of-distribution samples is effectively enhanced.

To further investigate the effectiveness of the margin-based contrastive loss in Equation (2), we replaced the margin constraint with a simple L2 loss for negative sample pairs, as shown in the seventh row. The results indicate a slight degradation across multiple metrics, which can be attributed to the fact that the L2 loss enforces unbounded repulsion for negative pairs that may disrupt the overall feature space, while the margin constraint introduces a controlled separation that better preserves intra-class compactness and inter-distribution margin.

Overall, the results demonstrate that each module is indispensable, and their synergy yields optimal performance.

Table 5. The importance of each component in FSCD. The experiments were conducted on DIGITS using LeNet-5. The best results are highlighted in bold, and the second-best results are underlined. * Indicates the use of the L2 loss in the contrastive learning module.

Pretraining	Fine-Tuning				AUROC	AUPR	FPR95	ACC
	DLPA		FCD					
	Image Level	Feature Level	Contrastive Learning	Denoising				
✓					49.61	81.05	93.23	47.01
	✓	✓	✓	✓	**83.37**	59.26	<u>74.57</u>	38.23
✓		✓	✓	✓	74.45	89.73	75.01	58.82
✓	✓		✓	✓	75.32	90.26	75.97	59.71
✓	✓	✓		✓	51.32	82.77	92.18	49.48
✓	✓	✓	✓		76.21	<u>91.84</u>	73.66	<u>60.13</u>
✓	✓	✓	*		75.94	90.21	75.20	59.34
✓	✓	✓	✓	✓	<u>76.49</u>	**92.34**	**73.43**	**60.44**

4.6. Impact of Backbone Networks

To assess the generality of FSCD, comparative experiments were conducted using different network models. The primary experiment employed the widely used ResNet-18 as the base architecture. In addition, DenseNet-100, which leverages dense connectivity to enhance information flow between layers, mitigate gradient vanishing, and improve parameter efficiency, was evaluated to further validate the effectiveness of FSCD.

Table 6 presents the experimental results on the OBJECTS benchmark. The results indicate that, under the more complex DenseNet-100 architecture, FSCD continued to exhibit performance advantages. FSCD achieves higher scores in key metrics such as AUROC and AUPR compared to the other methods in most cases. This further confirms the stability and superiority of FSCD across different network architectures, implying that it can adapt to various backbone networks without significant performance fluctuations due to architectural changes, thereby providing strong support for practical applications.

Table 6. Results under different architectures. The experiments were conducted on OBJECTS using ResNet-18 and DenseNet-100. The best results are highlighted in bold, and the second-best results are underlined.

Model	Method	AUROC	AUPR	FPR95
ResNet-18	MSP	72.03	84.04	88.70
	EBO	66.65	78.13	**81.54**
	MDS	73.75	84.73	82.33
	ViM	74.09	83.02	82.57
	SEM	<u>77.52</u>	<u>89.97</u>	88.84
	Ours	**78.07**	**90.15**	<u>82.65</u>
DenseNet-100	MSP	75.92	87.44	85.23
	EBO	67.60	82.91	<u>77.95</u>
	MDS	77.68	86.36	79.00
	ViM	75.60	86.81	82.04
	SEM	<u>79.85</u>	<u>92.07</u>	86.51
	Ours	**81.03**	**92.27**	**74.51**

4.7. Influence of Hyperparameters

This section examines three hyperparameters: λ, λ_1, and k. The parameter λ is used during the semantic boundary initialization phase to balance the classification loss of in-distribution samples and the loss from auxiliary outlier samples. As λ is used only for pretraining, it was not analyzed in the ablation study. In contrast, k in Equation (2) modulates the weight of the distances among perturbed samples, in-distribution samples, and auxiliary outliers, and λ_1 in Equation 3 adjusts the contribution of the classification loss relative to the contrastive denoising module. Ablation experiments on the DIGITS benchmark were performed to evaluate them. AUROC was used to assess the model's OOD detection capability, while classification accuracy reflects the performance on in-distribution samples. Both k and λ_1 were varied between 0.1 and 0.9 to explore performance trends and identify optimal settings.

As shown in Figure 3, the classification accuracy initially increased with the growth of k before subsequently decreasing, while AUROC peaked at $k = 0.3$ and then slightly declined and stabilized. This indicates that an appropriate value of k effectively distinguished ID samples, perturbed samples, and OOD samples by increasing their separation, thereby enhancing OOD detection performance. However, when k became excessively large, the overemphasis on the contrastive loss reduced the model's focus on differentiating between ID samples and perturbed samples. Consequently, further increases in k failed to yield significant performance gains and may have even impaired the classification capability.

As shown in Figure 4, when λ_1 was small, the influence of the contrastive denoising module was limited, and the model primarily enhanced ID classification without

sufficiently distinguishing samples with covariate shifts, leading to suboptimal overall performance. Similarly, Figure 4 indicates that the OOD detection capability was inadequate at low λ_1 values. As λ_1 increased, improvements in both classification accuracy and AUROC were observed, suggesting that an appropriate emphasis on the contrastive denoising module enhances the model's ability to differentiate among ID samples, perturbed samples, and outliers. However, when λ_1 exceeded 0.5, the classification accuracy began to decline, likely because the model became overly focused on contrastive learning at the expense of the primary classification task.

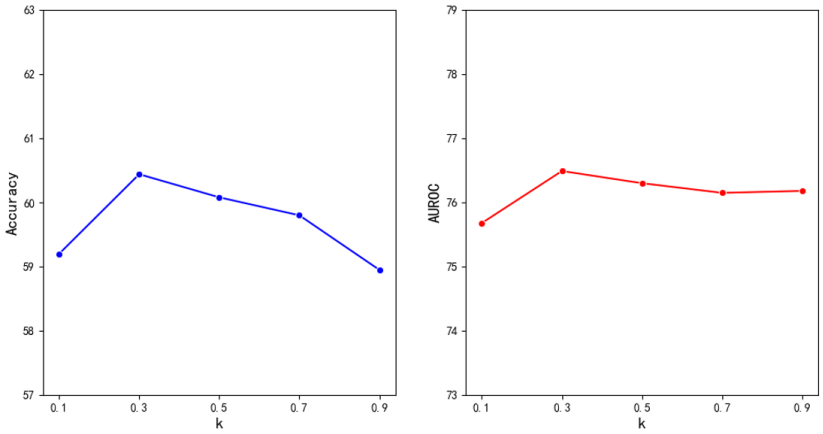


Figure 3. The results for accuracy (**left**) and AUROC (**right**) under different weight factors k. The experiments were conducted on DIGITS using LeNet-5.

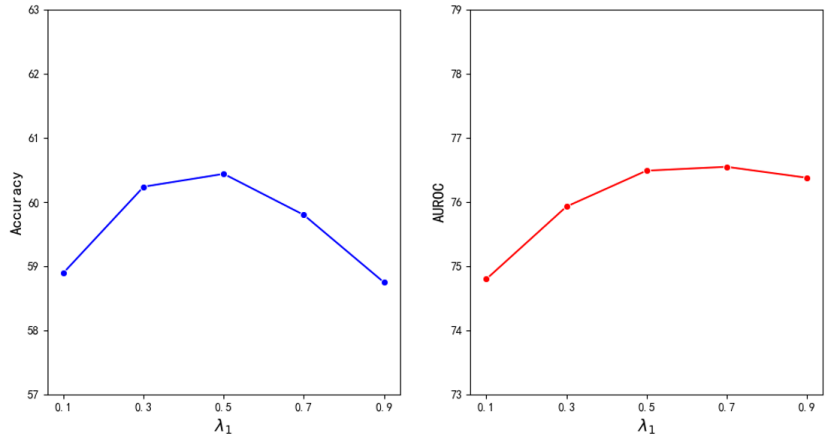


Figure 4. The results for accuracy (**left**) and AUROC (**right**) under different weight factors λ_1. The experiments were conducted on DIGITS using LeNet-5.

5. Limitations and Future Research Directions

The FSCD framework proposed in this paper achieved remarkable results in full-spectrum out-of-distribution detection. Despite its promising results on smaller datasets, the performance of the model on larger-scale datasets has yet to be evaluated, primarily due to the limitations of existing Fs-OOD benchmarks. Additionally, upon examining failure cases, we observed that FSCD had a higher miss rate for certain far-OOD samples. This finding indicates that our simulation of covariate-shift samples may have been excessively broad, necessitating a more refined approach.

In future research, two primary directions could be pursued. First, in real-world scenarios, data distributions are inherently dynamic, yet most current OOD detection methodologies are geared toward static distributions. Future work could explore the integration of incremental learning with OOD detection, allowing models to dynamically

adapt to real-time data updates and maintain robust detection performance in evolving environments. Second, in outlier exposure-based OOD detection, the quality of outlier samples plays a critical role in model performance. However, selecting high-quality outliers is often labor-intensive. Future research could focus on reducing dependence on auxiliary outlier samples and developing more universal and efficient OOD detection strategies.

6. Conclusions

This paper addresses the challenge of full-spectrum out-of-distribution detection and proposes the Full-Spectrum Contrastive Denoising (FSCD) framework. FSCD introduces a dual-stage training framework consisting of a semantic initialization phase and a fine-tuning phase. During the semantic initialization phase, FSCD establishes a reliable initial semantic classification boundary by leveraging both in-distribution and auxiliary outlier samples. In the fine-tuning phase, FSCD further enhances model robustness by employing a dual-level perturbation augmentation module to simulate covariate shift samples and a feature contrastive denoising module to increase the discriminative separation between OOD and ID samples.

Extensive experiments are conducted on three benchmarks—DIGITS, OBJECTS, and COVID—demonstrating the strong performance and generalizability of the proposed FSCD framework. It achieves consistent improvements across all benchmarks, outperforming existing methods in both detection accuracy and robustness. In particular, FSCD achieves the highest AUROC values of 76.49%, 78.07%, and 90.11%, respectively. Furthermore, it obtains the highest classification accuracy on covariate-shifted in-distribution samples, highlighting its effectiveness in handling real-world variations.

This work provides novel insights and methodologies for real-world OOD detection, offering both theoretical significance and practical value.

Author Contributions: Conceptualization, D.P.; methodology, D.P.; software, D.P.; validation, B.S. and X.L.; formal analysis, D.P. and B.S.; investigation, D.P. and B.S.; resources, B.S. and X.L.; data curation, D.P. and B.S.; writing—original draft preparation, D.P.; writing—review and editing, B.S. and X.L.; visualization, D.P.; supervision, B.S. and X.L.; project administration, B.S. and X.L.; funding acquisition, X.L. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the Science and Technology Innovation Plan of the Shanghai Science and Technology Commission under grant number 22511106005.

Data Availability Statement: Data will be made available on request.

Acknowledgments: The authors would like to thank the High-Performance Computing Center of Shanghai University and the Shanghai Engineering Research Center of Intelligent Computing Systems for providing the computing resources and technical support.

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. He, K.; Zhang, X.; Ren, S.; Sun, J. Deep residual learning for image recognition. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 27–30 June 2016; pp. 770–778.
2. Ren, S.; He, K.; Girshick, R.; Sun, J. Faster R-CNN: Towards real-time object detection with region proposal networks. *IEEE Trans. Pattern Anal. Mach. Intell.* **2016**, *39*, 1137–1149. [CrossRef] [PubMed]
3. Long, J.; Shelhamer, E.; Darrell, T. Fully convolutional networks for semantic segmentation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Boston, MA, USA, 7–12 June 2015; pp. 3431–3440.
4. Nguyen, A.; Yosinski, J.; Clune, J. Deep neural networks are easily fooled: High confidence predictions for unrecognizable images. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Boston, MA, USA, 7–12 June 2015; pp. 427–436.
5. Hendrycks, D.; Gimpel, K. A baseline for detecting misclassified and out-of-distribution examples in neural networks. In Proceedings of the International Conference on Learning Representations, Toulon, France, 24–26 April 2017; pp. 1–12.

6. Cui, P.; Wang, J. Out-of-distribution (OOD) detection based on deep learning: A review. *Electronics* **2022**, *11*, 3500. [CrossRef]
7. Zhang, J.; Yang, J.; Wang, P.; Wang, H.; Lin, Y.; Zhang, H.; Sun, Y.; Du, X.; Zhou, K.; Zhang, W.; et al. OpenOOD v1.5: Enhanced Benchmark for Out-of-Distribution Detection. In Proceedings of the NeurIPS 2023 Workshop on Distribution Shifts: New Frontiers with Foundation Models, New Orleans, LA, USA, 15 December 2023.
8. Yang, J.; Zhou, K.; Liu, Z. Full-spectrum out-of-distribution detection. *Int. J. Comput. Vis.* **2023**, *131*, 2607–2622. [CrossRef]
9. Hsu, Y.C.; Shen, Y.; Jin, H.; Kira, Z. Generalized odin: Detecting out-of-distribution image without learning from out-of-distribution data. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 13–19 June 2020; pp. 10951–10960.
10. Long, X.; Zhang, J.; Shan, S.; Chen, X. Rethinking the Evaluation of Out-of-Distribution Detection: A Sorites Paradox. *Adv. Neural Inf. Process. Syst.* **2024**, *37*, 89806–89833.
11. Averly, R.; Chao, W.L. Unified out-of-distribution detection: A model-specific perspective. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Paris, France. 2–3 October 2023; pp. 1453–1463.
12. Yang, W.; Zhang, B.; Russakovsky, O. Imagenet-Ood: Deciphering Modern Out-of-Distribution Detection Algorithms. In Proceedings of the 12th International Conference on Learning Representations, International Conference on Learning Representations 2024, Vienna Austria, 7–11 May 2024.
13. Yang, J.; Wang, P.; Zou, D.; Zhou, Z.; Ding, K.; Peng, W.; Wang, H.; Chen, G.; Li, B.; Sun, Y.; et al. Openood: Benchmarking generalized out-of-distribution detection. *Adv. Neural Inf. Process. Syst.* **2022**, *35*, 32598–32611.
14. Hendrycks, D.; Mazeika, M.; Dietterich, T. Deep Anomaly Detection with Outlier Exposure. *arXiv* **2018**, arXiv:1812.04606.
15. Papadopoulos, A.A.; Rajati, M.R.; Shaikh, N.; Wang, J. Outlier exposure with confidence control for out-of-distribution detection. *Neurocomputing* **2021**, *441*, 138–150. [CrossRef]
16. Zhang, J.; Inkawhich, N.; Linderman, R.; Chen, Y.; Li, H. Mixture outlier exposure: Towards out-of-distribution detection in fine-grained environments. In Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, Paris, France, 2–3 October 2023; pp. 5531–5540.
17. Choi, H.; Jeong, H.; Choi, J.Y. Balanced energy regularization loss for out-of-distribution detection. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Paris, France, 2–3 October 2023; pp. 15691–15700.
18. Xu, K.; Chen, R.; Franchi, G.; Yao, A. Scaling for Training Time and Post-hoc Out-of-distribution Detection Enhancement. In Proceedings of the International Conference on Learning Representations, Vienna Austria, 7–11 May 2024.
19. Liang, S.; Li, Y.; Srikant, R. Enhancing the reliability of out-of-distribution image detection in neural networks. In Proceedings of the International Conference on Learning Representations, Vancouver, BC, Canada, 30 April–3 May 2018; pp. 1–12.
20. Liu, W.; Wang, X.; Owens, J.; Li, Y. Energy-based out-of-distribution detection. *Adv. Neural Inf. Process. Syst.* **2020**, *33*, 21464–21475.
21. Sun, Y.; Guo, C.; Li, Y. React: Out-of-distribution detection with rectified activations. *Adv. Neural Inf. Process. Syst.* **2021**, *34*, 144–157.
22. Hendrycks, D.; Basart, S.; Mazeika, M.; Zou, A.; Kwon, J.; Mostajabi, M.; Steinhardt, J.; Song, D. Scaling Out-of-Distribution Detection for Real-World Settings. In Proceedings of the International Conference on Machine Learning, PMLR, Baltimore, MD, USA, 17–23 July 2022; pp. 8759–8773.
23. Krumpl, G.; Avenhaus, H.; Possegger, H.; Bischof, H. Ats: Adaptive temperature scaling for enhancing out-of-distribution detection methods. In Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, Waikoloa, HI, USA, 1–6 January 2024; pp. 3864–3873.
24. Karunanayake, N.; Seneviratne, S.; Chawla, S. ExCeL: Combined Extreme and Collective Logit Information for Out-of-Distribution Detection. *arXiv* **2023**, arXiv:2311.14754.
25. Zhang, S.; Zhang, X.; Wan, S.; Ren, W.; Zhao, L.; Shen, L. Generative adversarial and self-supervised dehazing network. *IEEE Trans. Ind. Inform.* **2023**, *20*, 4187–4197. [CrossRef]
26. Liu, Y.; Wang, X.; Hu, E.; Wang, A.; Shiri, B.; Lin, W. VNDHR: Variational Single Nighttime Image Dehazing for Enhancing Visibility in Intelligent Transportation Systems via Hybrid Regularization. *IEEE Trans. Intell. Transp. Syst.* **2025**, *2025*, 1–15. [CrossRef]
27. Tack, J.; Mo, S.; Jeong, J.; Shin, J. Csi: Novelty detection via contrastive learning on distributionally shifted instances. *Adv. Neural Inf. Process. Syst.* **2020**, *33*, 11839–11852.
28. Gwon, K.; Yoo, J. Out-of-distribution (OOD) detection and generalization improved by augmenting adversarial mixup samples. *Electronics* **2023**, *12*, 1421. [CrossRef]
29. Wei, T.; Wang, B.L.; Zhang, M.L. EAT: Towards Long-Tailed Out-of-Distribution Detection. In Proceedings of the AAAI Conference on Artificial Intelligence, Vancouver, BC, Canada, 20–27 February 2024; Volume 38, pp. 15787–15795.
30. Feng, S.; Wang, C. When an extra rejection class meets out-of-distribution detection in long-tailed image classification. *Neural Netw.* **2024**, *178*, 106485. [CrossRef]
31. Sohn, K.; Berthelot, D.; Carlini, N.; Zhang, Z.; Zhang, H.; Raffel, C.A.; Cubuk, E.D.; Kurakin, A.; Li, C.L. Fixmatch: Simplifying semi-supervised learning with consistency and confidence. *Adv. Neural Inf. Process. Syst.* **2020**, *33*, 596–608.

32. Yang, L.; Qi, L.; Feng, L.; Zhang, W.; Shi, Y. Revisiting weak-to-strong consistency in semi-supervised semantic segmentation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Vancouver, BC, Canada, 17–24 June 2023; pp. 7236–7246.
33. Zhao, H.; Gallo, O.; Frosio, I.; Kautz, J. Loss functions for image restoration with neural networks. *IEEE Trans. Comput. Imaging* **2016**, *3*, 47–57. [CrossRef]
34. LeCun, Y.; Bottou, L.; Bengio, Y.; Haffner, P. Gradient-based learning applied to document recognition. *Proc. IEEE* **1998**, *86*, 2278–2324. [CrossRef]
35. Hull, J.J. A database for handwritten text recognition research. *IEEE Trans. Pattern Anal. Mach. Intell.* **1994**, *16*, 550–554. [CrossRef]
36. Xiao, H. Fashion-mnist: A novel image dataset for benchmarking machine learning algorithms. *arXiv* **2017**, arXiv:1708.07747.
37. Cimpoi, M.; Maji, S.; Kokkinos, I.; Mohamed, S.; Vedaldi, A. Describing textures in the wild. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Columbus, OH, USA, 23–28 June 2014; pp. 3606–3613.
38. Cao, K.; Wei, C.; Gaidon, A.; Arechiga, N.; Ma, T. Learning imbalanced datasets with label-distribution-aware margin loss. *Adv. Neural Inf. Process. Syst.* **2019**, *32*.
39. Le, Y.; Yang, X. Tiny imagenet visual recognition challenge. *CS 231N* **2015**, *7*, 3.
40. Zhou, B.; Lapedriza, A.; Khosla, A.; Oliva, A.; Torralba, A. Places: A 10 million image database for scene recognition. *IEEE Trans. Pattern Anal. Mach. Intell.* **2017**, *40*, 1452–1464. [CrossRef] [PubMed]
41. Russakovsky, O.; Deng, J.; Su, H.; Krause, J.; Satheesh, S.; Ma, S.; Huang, Z.; Karpathy, A.; Khosla, A.; Bernstein, M.; et al. Imagenet large scale visual recognition challenge. *Int. J. Comput. Vis.* **2015**, *115*, 211–252. [CrossRef]
42. Santa Cruz, B.G.; Bossa, M.N.; Sölter, J.; Husch, A.D. Public COVID-19 x-ray datasets and their impact on model bias–a systematic review of a significant problem. *Med. Image Anal.* **2021**, *74*, 102225. [CrossRef]
43. Wang, L.; Lin, Z.Q.; Wong, A. Covid-net: A tailored deep convolutional neural network design for detection of covid-19 cases from chest x-ray images. *Sci. Rep.* **2020**, *10*, 19549. [CrossRef]
44. Winther, H.B.; Laser, H.; Gerbel, S.; Maschke, S.K.; Hinrichs, J.B.; Vogel-Claussen, J.; Wacker, F.K.; Höper, M.M.; Meyer, B.C. COVID-19 Image Repository. 2020. Available online: https://figshare.com/articles/dataset/COVID-19_Image_Repository/12275009 (accessed on 8 February 2025). [CrossRef]
45. Yang, X.; He, X.; Zhao, J.; Zhang, Y.; Zhang, S.; Xie, P. Covid-ct-dataset: A ct scan dataset about COVID-19. *arXiv* **2020**, arXiv:2003.13865.
46. RSNA. *RSNA Pediatric Bone Age Challenge*; RSNA: Oak Brook, IL, USA, 2017.
47. Torralba, A.; Fergus, R.; Freeman, W.T. 80 million tiny images: A large data set for nonparametric object and scene recognition. *IEEE Trans. Pattern Anal. Mach. Intell.* **2008**, *30*, 1958–1970. [CrossRef]
48. Lee, K.; Lee, K.; Lee, H.; Shin, J. A simple unified framework for detecting out-of-distribution samples and adversarial attacks. *Adv. Neural Inf. Process. Syst.* **2018**, *31*, 1–11.
49. Wang, H.; Li, Z.; Feng, L.; Zhang, W. Vim: Out-of-distribution with virtual-logit matching. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, New Orleans, LA, USA, 18–24 June 2022; pp. 4921–4930.
50. Chen, Y.; Wu, Y.; Wang, T.; Zhang, S.; Zhu, D.; Liu, C. Mitigating covariance overfitting in out-of-distribution detection through intrinsic parameter learning. *Knowl.-Based Syst.* **2025**, *316*, 113318. [CrossRef]

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.