FROB: Few-shot ROBust Model for Joint Classification and Out-of-Distribution Detection

Nikolaos Dionelis[1 [0000−0001−9662−8537]] ✉, Sotirios A. Tsaftaris[1 [0000−0002−8795−9294]], and Mehrdad Yaghoobi[1 [0000−0002−9847−8234]]

[1] The University of Edinburgh, UK
School of Engineering, Electrical Engineering, Digital Communications
Contact email: nikolaos.dionelis@ed.ac.uk

Abstract. Classification and Out-of-Distribution (OoD) detection in the few-shot setting remain challenging aims, but are important for devising critical systems in security where samples are limited. OoD detection requires that classifiers are aware of when they do not know and avoid setting high confidence to OoD samples away from the training data distribution. To address such limitations, we propose the Few-shot ROBust (FROB) model with its key contributions being (a) the joint classification and few-shot OoD detection, (b) the sample generation on the boundary of the support of the normal class distribution, and (c) the incorporation of the learned distribution boundary as OoD data for contrastive negative training. FROB finds the boundary of the support of the normal class distribution, and uses it to improve the few-shot OoD detection performance. We propose a self-supervised learning methodology for sample generation on the normal class distribution confidence boundary based on generative and discriminative models, including classification. FROB implicitly generates adversarial samples, and forces samples from OoD, including our boundary, to be less confident by the classifier. By including the learned boundary, FROB reduces the threshold linked to the model's few-shot robustness in the number of few-shots, and maintains the OoD performance approximately constant, independent of the number of few-shots. The low- and few-shot robustness evaluation of FROB on different image datasets and on One-Class Classification (OCC) data shows that FROB achieves competitive performance and outperforms baselines in terms of robustness to the OoD few-shot population and variability.

Keywords: Out-of-Distribution detection · Few-shot anomaly detection

1 Introduction

In real-world settings, for AI-enabled systems to be operational, it is crucial to robustly perform joint classification and Out-of-Distribution (OoD) detection, and report an input as OoD rather than misclassifying it. The problem of detecting whether a sample is in-distribution, i.e. from the training distribution, or OoD is critical. This is crucial in safety and security as the consequences of failure to detect OoD objects can be severe and eventually fatal. However, deep

neural networks produce overconfident predictions and do not distinguish in- and out-of-data-distribution. Adversarial examples, when small modifications of the input appear, can change the classifier decision. It is an important property of a classifier to address such limitations and provide robustness guarantees. In parallel, OoD detection is challenging as classifiers set high confidence to OoD samples away from the training data. In this paper, we propose the Few-shot ROBust (FROB) model to accurately perform simultaneous classification and OoD detection in the few-shot setting. To address rarity and the existence of limited samples in the few-shot setting [1,2], we aim at reducing the number of few-shots of OoD data required, whilst maintaining accurate and robust performance.

Training with outlier sets of diverse data, available today in large quantities, can improve OoD detection [3,4,5]. General OoD datasets enable OoD generalization to detect unseen OoD with improved robustness and performance. Models trained with different outliers can detect OoD by learning cues for whether inputs lie within or out of the support of the normal class distribution. By exposing models to different OoD, the complement of the support of the normal class distribution is modelled. The detection of new types of OoD is enabled. OoD datasets improve the calibration of classifiers in the setting where a fraction of the data is OoD, addressing overconfidence issues when applied to OoD [3,4].

The main benefits of FROB are that (a) joint classification and OoD detection is realistic, effective, and beneficial, (b) our proposed distribution boundary is a principled, effective, and beneficial approach to generate near OoD samples for negative training, and (c) contrastive training to include the learned negative data during training is effective and beneficial. Furthermore, the benefits of performing joint multi-class classification and OoD detection are that (i) this setting is more realistic and has wider applicability because in the real-world, models should be both operational and reliable and declare an input as OoD rather than misclassifying it, (ii) using discriminative classifier models leads to improved OoD detection performance, and (iii) in the few-shot setting, discriminative classifiers address the limited data problem with improved robustness. An additional benefit of performing simultaneous classification and OoD detection is that we take advantage of labelled data to achieve improved anomaly detection performance as they contain more information because of their labels and classes. Knowing the normal data better, as well as learning how the data are structured in clusters with class labels, helps us to detect OoD data better.

We address the rarity of near and relevant anomalies during training by performing sample generation on the boundary of the support of the underlying distribution of the data from the normal class. The benefit of this is improved robustness to the OoD few-shot population and variability. Task-specific OoD samples are hard to find in practice; in the real world, we also have budget limitations for (negative) sampling. FROB achieves significantly better robustness for few-shot OoD detection, while maintaining in-distribution accuracy. Aiming at solving the few-shot robustness problem with classification and OoD detection, the contribution of our FROB methodology is the development of an integrated robust framework for self-supervised few-shot negative data augmentation on

the distribution confidence boundary, combined with few-shot OoD detection. FROB trains a generator to create low-confidence samples on the normal class boundary, and includes these learned samples in the training to improve the performance in the few-shot setting. The combination of the self-generated boundary and the imposition of low confidence at this learned boundary is a contribution of FROB, which improves robustness for few-shot OoD detection. The main benefits of our distribution boundary framework are that it is a principled approach based on distributions, it generates near-OoD samples that are well-sampled and evenly scattered, these near negative data are strong anomalies and adversarial anomalies [11,6], and these learned OoD data are the closest possible negative samples to the normal class. This latter characteristic of our algorithm leads to the tightest-possible OoD data description and characterization, and to self-generated negatives that are optimal in the sense that no unfilled space is allowed between the normal class data and the *learned* OoD samples. In this way, FROB uses the definition of anomaly and the delimitation of the support boundary of the normal class distribution, which are needed for improved robustness.

We achieve generalization to unseen OoD, with applicability to new unknown, in the wild, test sets that do not correlate to the training sets. FROB's evaluation in several settings, using cross-dataset and One-Class Classification (OCC) evaluations, shows that key methodological contributions such as generating samples on the normal class distribution boundary and few-shot adaptation, improve few-shot OoD detection. Our experiments show robustness to the number of OoD few-shots and to outlier variation, outperforming methods we compare with.

2 Proposed Methodology for Few-Shot OoD Detection

We propose FROB in Fig. 1 for joint classification and few-shot OoD detection combining discriminative and generative models. We aim for improved robustness and reliable confidence prediction, and force low confidence close and away from the data. Our key idea is to jointly learn a classifier but also a generative model that finds the boundary of the support of the in-distribution data. We use this generator to create adversarial samples on the boundary close to the in-distribution data. We combine these in a self-supervised learning manner, where the generated data act as a negative class. We propose a robustness loss to classify as less confident samples on, and out of, the learned boundary. FROB also uses few-shots of real OoD data naturally within the formulation we propose.

Loss function. We denote the normal class data by $\mathbf{x}$ where $\mathbf{x}_i$ are the labeled data, with labels y_i between 1 and K. The few-shot OoD samples are $\mathbf{Z}_m$. The cost function of the classifier model, minimized during training, is

$$\begin{aligned} \arg\min_f \; & -\frac{1}{N}\sum_{i=1}^{N}\log\frac{\exp(f_{y_i}(\mathbf{x}_i))}{\sum_{k=1}^{K}\exp(f_k(\mathbf{x}_i))} \\ & -\lambda\frac{1}{M}\sum_{m=1}^{M}\log\left(1-\max_{l=1,2,\ldots,K}\frac{\exp(f_l(\mathbf{Z}_m))}{\sum_{k=1}^{K}\exp(f_k(\mathbf{Z}_m))}\right), \end{aligned} \tag{1}$$

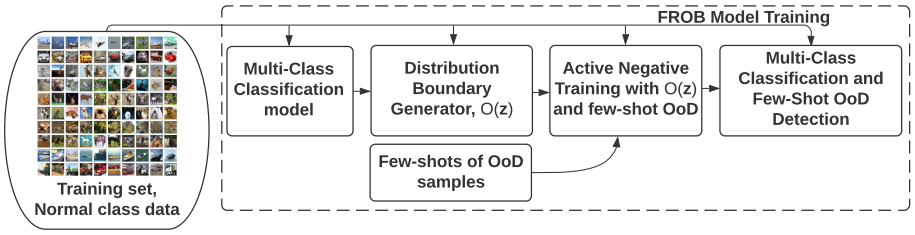

Fig. 1: FROB training with learned negative sampling, $O(\mathbf{z})$, and few-shot OoD.

where $f(\cdot)$ is the Convolutional Neural Network (CNN) discriminative model for multi-class classification with K classes. The proposed objective cost function has two loss terms and a hyperparameter. The two loss terms operate on different samples for positive and negative training, respectively. The first loss term is the cross-entropy between y_i and the predictions, $\text{softmax}(f(\mathbf{x}_i))$; the CNN is followed by the normalized exponential to obtain the probability over the classes. The second loss term enforces $f(\cdot)$ to more accurately detect outliers, in addition to performing multi-class classification. It is weighted by a hyperparameter, λ.

FROB then trains a generator to generate low-confidence samples on the normal class distribution boundary. Our algorithm includes these learned low-confidence samples in the training to improve the performance in the few-shot setting. We do not use a large general OoD dataset because general-purpose OoD datasets lead to an ad hoc selection of outliers that try to approximate data outside the support of the normal class distribution. Instead, we use negative data augmentation and self-supervised learning to model the boundary of the support of the normal class distribution. Our proposed FROB model generates outliers via a trained generator $O(\mathbf{z})$, which takes the form of a CNN. Here, O refers to OoD samples, and $\mathbf{z}$ are samples from a standard Gaussian distribution. The optimization of maximizing dispersion subject to being on the boundary is

$$\begin{aligned} \arg \ \min_O \ \frac{1}{N-1} \sum_{j=1,\, \mathbf{z}_j \neq \mathbf{z}}^{N} \frac{||\mathbf{z}-\mathbf{z}_j||_2}{||O(\mathbf{z})-O(\mathbf{z}_j)||_2} \ + \ \nu \min_{j=1,2,\ldots,Q} ||O(\mathbf{z})-\mathbf{x}_j||_2 \\ + \ \mu \max_{l=1,2,\ldots,K} \frac{\exp(f_l(O(\mathbf{z}))-f_l(\mathbf{x}))}{\sum_{k=1}^{K} \exp(f_k(O(\mathbf{z}))-f_k(\mathbf{x}))}, \end{aligned} \tag{2}$$

where by using (2), we penalize the probability that $O(\mathbf{z})$ has higher confidence than the normal class. We make $O(\mathbf{z})$ have lower confidence than $\mathbf{x}$. FROB includes the learned low-confidence samples in the training by performing (1) with the self-generated boundary, $O(\mathbf{z})$, instead of $\mathbf{Z}$. Our self-supervised learning mechanism to calibrate prediction confidence in unforeseen scenarios is (2) followed by (1). We perform distribution boundary data augmentation in a learnable manner and set this distribution confidence boundary as negative data to improve few-shot OoD detection. This learned boundary includes strong and specifically adversarial anomalies close to the distribution support and near high

probability normal class data. FROB sets samples just outside the data distribution boundary as OoD. We introduce relevant anomalies to more accurately and more robustly detect few-shots of OoD [2]. We detect OoD samples by generating task-specific anomalous samples. We employ a nested optimization: an inner optimization to find $O(\mathbf{z})$ in (2), and an outer optimization based on cross-entropy with negative training in (1). For this nested optimization, if an optimum point is reached for the inner one, an optimum will also be reached for the outer.

FROB, for robust OoD detection, performs negative data augmentation on the support boundary of the normal class in a well-sampled manner. Specifically, FROB performs OoD sample description and characterization. By using (2), it does not allow for unfilled space between the normal class and the self-generated OoD. The second loss term of our loss function in (2) is designed to not permit unused and slack space between the learned negatives and the normal class data [11,6]. Our learned near-OoD samples have low point-to-set distance as measured by the second loss term of our proposed objective cost function shown in (2).

In the proposed self-supervised approach, the loss function for the parameter updation in the generator is (2). The first loss term is for scattering the generated samples. This measure reduces mode collapse and preserves distance proportionality in the latent and data spaces. The second loss term penalizes deviations from normality by using the distance from a point to a set. The third term in (2) guides to find the data distribution boundary by penalizing prediction confidence and pushing the generated samples OoD. In the second term, we denote the data by $(\mathbf{x}_j, y_j)_{j=1}^{Q}$, e.g. $\mathbf{x}_j$ is a vector of length 3072 for CIFAR-10.

By employing (2) followed by (1), FROB addresses the question of what OoD samples to introduce to our model for negative training in order to accurately and robustly detect few-shot data and achieve good few-shot generalization. FROB introduces self-supervised learning and learned negative data augmentation using the tightest-possible OoD data description algorithm of (2) followed by (1). Our distribution confidence boundary in (2) is robust to the problem of generators not capturing the entire data distribution and eventually learning only a Dirac distribution, which is known as mode collapse [6,7]. Using scattering, we achieve sample diversity by using the ratio of distances in the latent and data spaces. In addition, in (2), our FROB model also uses data space point-set distances.

FROB redesigns and streamlines the use of general OoD datasets to work for few-shot samples, even for zero-shots, using self-supervised learning to model the boundary of the support of the normal class distribution instead of using a large OoD set. Such general-purpose large OoD sets lead to an ad hoc selection of outliers trying to model the complement of the support of the normal class distribution. The boundary of the support of the normal class distribution, which FROB finds using (2), has and needs *less samples* than the entire complement of the support of the data distribution that big OoD sets try to approximate.

Inference. The Anomaly Score (AS) of FROB for *any* queried test sample, $\tilde{\mathbf{x}}$, in the data space, during inference and model deployment, is given by

$$\mathrm{AS}(f, \tilde{\mathbf{x}}) \;=\; \max_{l=1,2,\ldots,K} \frac{\exp(f_l(\tilde{\mathbf{x}}))}{\sum_{k=1}^{K} \exp(f_k(\tilde{\mathbf{x}}))}, \tag{3}$$

where l is the decided class. If the AS of $\tilde{\mathbf{x}}$ has a value smaller than a predefined threshold, τ, i.e. AS $< \tau$, then $\tilde{\mathbf{x}}$ is OoD. Otherwise, $\tilde{\mathbf{x}}$ is in-distribution data.

3 Related Work on Classification and OoD Detection

General OoD datasets. Training detectors using outliers from general OoD datasets can improve the OoD detection performance to detect unseen anomalies [5]. Using datasets disjoint from train and test data, models can learn representations for OoD detection. Confidence Enhancing Data Augmentation (CEDA), Adversarial Confidence Enhancing Training (ACET), and Guaranteed OoD Detection (GOOD) address the overconfidence of classifiers at OoD samples [3,4]. They enforce low confidence in a l_∞-norm ball around each OoD sample. CEDA employs point-wise robustness [13]. GOOD finds worst-case OoD detection guarantees. The models are trained on general OoD datasets that are, however, reduced by the normal class dataset. Disjoint distributions are used for positive and negative training, but the OoD samples are selected in an ad hoc manner. In contrast, FROB performs learned negative data augmentation on the normal class distribution confidence boundary to redesign few-shot OoD detection.

Human prior. GOOD first defines the normal class, and then filters it out from the general-purpose OoD dataset. This filtering-out process of normality from the general OoD dataset is human-dependent. It is not practical and cannot be used in the real world as anomalies are not confined to a finite labelled closed set [15]. This modified dataset is set as anomalies. Next, GOOD learns the normal class, and sets low confidence to these OoD. This process is not automatic and data- and feature-dependent [10,11]. In contrast, FROB eliminates the need for feature extraction and human intervention which is the aim of deep learning, as they do not scale. FROB avoids application and dataset dependent processes.

Learned negatives. The Confidence-Calibrated Classifier (CCC) uses a GAN to create samples out of, but also close to, the normal class distribution [9]. FROB substantially differs from CCC that finds a threshold and not the normal class distribution boundary. CCC uses a general OoD dataset, $U(\mathbf{y})$, where the labels follow a Uniform distribution, to compute this threshold. This can be limiting as the threshold depends on $U(\mathbf{y})$, which is an ad hoc selection of outliers that are located randomly somewhere in the data space. This leads to unfilled space between the OoD samples and the normal class which is suboptimal. In contrast, FROB finds the normal class distribution boundary and does not use $U(\mathbf{y})$ to find this boundary. Our distribution boundary is not a function of $U(\mathbf{y})$, as $U(\mathbf{y})$ is not necessary. For negative training, CCC defines a closeness metric (KL divergence), and penalizes it [11]. CCC suffers from mode collapse as it does not perform scattering for diversity. Confidence-aware classification is also performed in [9]. Self-Supervised outlier Detection (SSD) creates OoD samples in the Mahalanobis metric [8]. It is not a classifier, and it performs OoD detection with few-shot outliers. FROB achieves fast inference with (3), in contrast to [16] which is slow during inference. [16] does not address issues arising from detecting using nearest neighbours, while using a different composite loss for training.

4 Evaluation and Results

We evaluate FROB trained on different image datasets. For the evaluation of FROB, we examine the impact of different combinations of normal class datasets, OoD few-shots, and test datasets, in an alternating manner. We examine the generalization performance to few-shots of unseen OoD samples at the dataset level (out-of-dataset anomalies), which are different from the training sets.

Metrics. We report the Area Under the Receiver Operating Characteristic Curve (AUROC), Adversarial AUROC (AAUROC), and Guaranteed AUROC (GAUROC) [3,14]. To strengthen the robustness evaluation of FROB and to compare with benchmarks, in addition to AUROC, we also evaluate FROB with AAUROC and GAUROC. AAUROC and GAUROC are suitable for evaluating the robustness of OoD detection models focusing on the worst-case OoD detection performance using l_∞-norm perturbations for each of the OoD image samples. It uses the maximum confidence in the l_∞-norm ball around each OoD and finds a lower (upper respectively) bound on this maximum confidence. These worst-case confidences for the OoD samples are then used for the AUROC.

To examine the robustness to the number of few-shots, we decrease the number of OoD few-shots by dividing them by two, employing uniform sampling. We examine the influence of this on AUROC. Specifically, we examine the variation of AUROC, AAUROC, and GAUROC, which constitute the dependent variables, to changes of the independent variable, which is the provided number of few-shots of OoD samples. We examine the Breaking Point of our FROB algorithm and of benchmarks; we define this point as the number of few-shot data from which the OoD performance in AUROC decreases and then falls to 0.5.

Datasets. For the normal class, we use either CIFAR-10 or SVHN. For OoD few-shots, we use data from CIFAR-10, SVHN, CIFAR-100, and Low-Frequency Noise (LFN). To compare with baselines from the literature, for the general OoD datasets, we use SVHN, CIFAR-100, and the same general OoD dataset as in [3,5] but debiased, as in [18]. We evaluate our FROB model on the datasets CIFAR-100, SVHN, and CIFAR-10, as well as on LFN and Uniform noise.

Model architecture. FROB uses a CNN discriminative model, as described in Section 2. We also train and use a generator that takes the form of a CNN. We implement FROB in PyTorch and use the optimizer Adam for training.

Baselines. We demonstrate that FROB is effective and outperforms baselines in the few-shot OoD detection setting. We compare FROB to the baselines GEOM, GOAD, DROCC, Hierarchical Transformation Discriminating Generator (HTD), Support Vector Data Description (SVDD), and Patch SVDD (PSVDD) in the few-shot setting, using OCC [1]. We also compare FROB to GOOD [3], CEDA, CCC, OE and ACET [4], and [5]. For many-samples OoD, [3,5] use a general OoD set, which is not representative of the few-shot OoD detection setting. General OoD sets result in a nonoptimal ad hoc selection of OoD, especially when operating on a fixed few-shot budget for sampling from the OoD class.

Ablations. We evaluate FROB for few-shot OoD detection with (✓) the learned distribution boundary, $O(\mathbf{z})$, i.e. FROB. For ablation, we also evaluate models that are trained without (−) $O(\mathbf{z})$ samples which we term FROBInit.

Table 1: OoD performance of FROB with the learned distribution boundary, $O(\mathbf{z})$, in AUROC using OCC and few-shots of 80 CIFAR-10 anomalies, and comparison to baselines, [1]. *FODS is FROB with the outlier OoD dataset SVHN.*

NORMAL	DROCC	GEOM	GOAD	HTD	SVDD	PSVDD	FROB	FODS
Plane	0.790	0.699	0.521	0.748	0.609	0.340	**0.811**	**0.867**
Car	0.432	0.853	0.592	**0.880**	0.601	0.638	0.862	0.861
Bird	0.682	0.608	0.507	0.624	0.446	0.400	**0.721**	0.707
Cat	0.557	0.629	0.538	0.601	0.587	0.549	**0.748**	**0.787**
Deer	0.572	0.563	0.627	0.501	0.563	0.500	**0.742**	0.727
Dog	0.644	0.765	0.525	**0.784**	0.609	0.482	0.771	0.782
Frog	0.509	0.699	0.515	0.753	0.585	0.570	**0.826**	**0.884**
Horse	0.476	0.799	0.521	**0.823**	0.609	0.567	0.792	0.815
Ship	0.770	0.840	0.704	**0.874**	0.748	0.440	0.826	0.792
Truck	0.424	**0.834**	0.697	0.812	0.721	0.612	0.744	0.799
MEAN	0.585	0.735	0.562	0.756	0.608	0.510	**0.784**	**0.802**

4.1 Evaluation of FROB

Evaluation of FROB using OCC Compared to Baselines. We evaluate FROB using OCC for each CIFAR-10 class against several benchmarks in the few-shot setting of 80 samples [1]. FROB outperforms baselines in Table 1 which shows the mean performance of FROB when the normal class is a CIFAR-10 class. We compare our proposed FROB model to the baselines DROCC, GEOM, GOAD, HTD, SVDD, and PSVDD [1]. FROB with the self-learned $O(\mathbf{z})$ outperforms baselines for few-shot OoD detection in OCC when we have budget constraints and OoD sampling complexity limitations. We also evaluate our FROB model further retrained with the outlier OoD dataset SVHN, FODS, and show that using the OoD set is beneficial for few-shot OoD detection using OCC.

Robustness of FROB to the number of few-shots. We evaluate FROB with few-shots of OoD samples from SVHN in decreasing number, setting the normal class as CIFAR-10. We experimentally demonstrate the effectiveness of FROB, and the results are shown in Table 2 and Fig. 2. We evaluate FROB on SVHN, as well as on CIFAR-100 and LFN, in Fig. 2 where the in-distribution data are from the CIFAR-10 dataset while the OoD are from SVHN, CIFAR-100, and LFN. Using FROB, the performance improves showing robustness even for a small number of OoD few-shots, pushing down the phase transition point in the number of few-shots in Fig. 2. When the few-shots are from the test set, i.e. SVHN in Fig. 2, FROB is effective and robust for few-shot OoD detection.

Table 2: OoD performance of FROB using the learned boundary, $O(\mathbf{z})$, and OoD few-shots, tested on SVHN. The normal class is CIFAR-10 (C10). The second column shows the training data of *OoD few-shots* and their number.

MODEL	OoD FEW-SHOTS	TEST SET	AUROC	AAUROC	GAUROC
FROB	SVHN: 1830	SVHN	0.997	0.997	0.990
FROB	SVHN: 915	SVHN	0.995	0.995	0.984
FROB	SVHN: 732	SVHN	0.995	0.995	0.981
FROB	SVHN: 457	SVHN	0.997	0.997	0.982
FROB	SVHN: 100	SVHN	**0.996**	**0.996**	**0.950**
FROB	SVHN: 80	SVHN	**0.995**	**0.995**	**0.928**

We experimentally demonstrate that first performing sample generation on the distribution boundary, $O(\mathbf{z})$, and then *including* these learned OoD samples in our training is beneficial. The improvement of FROB in AUROC is because of these well-sampled $O(\mathbf{z})$ samples. The component of FROB with the highest benefit is the self-generated distribution boundary, $O(\mathbf{z})$. Our proposed FROB model shows improved robustness to the number of OoD few-shots because with decreasing few-shots, the performance of FROB in AUROC is robust and approximately independent of the OoD few-shot number of samples in Fig. 2.

Performance on Unseen Datasets. We evaluate FROB on OoD samples from unseen, in the wild, datasets, i.e. on samples that are neither from the normal class nor from the OoD few-shots. We examine our proposed FROB model in the few-shot setting in Fig. 2 for normal CIFAR-10 with OoD few-shots from SVHN, and tested on the new CIFAR-100 and LFN. These are unseen as they are not the normal class or the OoD few-shots. The performance of FROB in this OoD few-shot setting in Fig. 2 is robust on CIFAR-100 and on LFN.

Next, exchanging the datasets, FROB with the normal class SVHN, and a variable number of CIFAR-10 OoD few-shots, is tested in Table 3 and in Fig. 3. In Table 3, compared to Table 2, FROB achieves comparable performance for normal class SVHN and few-shots of CIFAR-10, compared to for normal class CIFAR-10 and few-shots of SVHN, in all the AUC-type metrics. According to Fig. 3, when compared to Fig. 2, for the unseen test set CIFAR-100, FROB achieves better AUROC for normal SVHN compared to for normal CIFAR-10.

Effect of domain and normal class. The performance of FROB in AUROC depends on the normal class. In Fig. 3, the OoD detection performance of FROB for small number of few-shots is higher for normal class SVHN than for normal CIFAR-10 in Fig. 2. FROB is robust and effective for normal SVHN on seen and unseen data. FROB is not sensitive to the number of few-shots for few-shot OoD detection, when we have OoD sample complexity constraints.

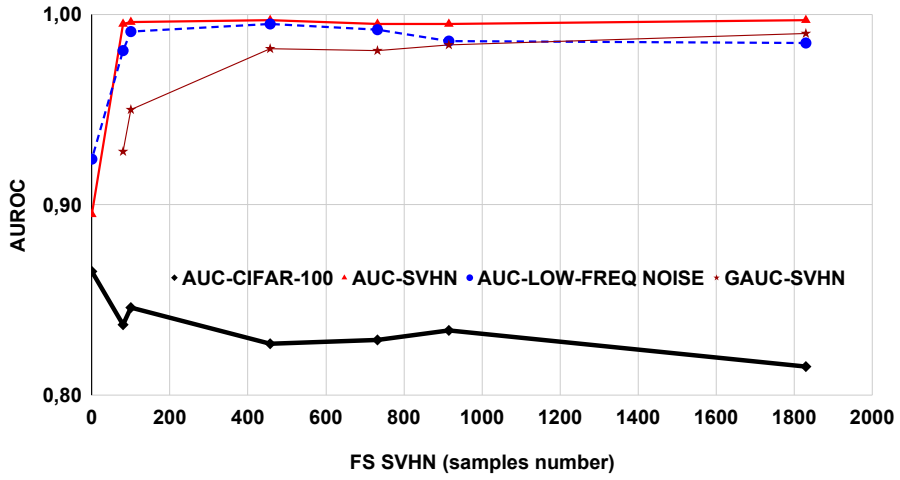

Fig. 2: FROB (normal C10) with SVHN OoD few-shots: AUROC and GAUROC.

Table 3: Evaluation of FROB for normal SVHN with the self-generated $O(\mathbf{z})$ and OoD few-shots, tested on CIFAR-10 (C10). According to the second and third columns, the OoD few-shots and the OoD test samples are from C10.

MODEL	OoD FEW-SHOTS	TEST SET	AUROC	AAUROC	GAUROC
FROB	C10: 600	C10	0.996	0.996	0.982
FROB	C10: 400	C10	**0.994**	**0.994**	**0.964**
FROB	C10: 200	C10	**0.996**	**0.996**	**0.967**
FROB	C10: 80	C10	**0.991**	**0.991**	**0.951**

OoD detection performance of FROB with OoD few-shots from the test set. In Tables 2 and 3 we experimentally demonstrate that FROB improves the AUROC and AAUROC when the few-shots and the test samples originate from the same dataset. We also show that FROB achieves high GAUROC.

OoD detection performance of FROB, OoD few-shots and test are different sets. More empirical results in Figs. 2 and 3 show that FROB also improves the AUROC when the few-shots and OoD test samples originate from different sets, i.e. LFN and CIFAR-100. This shows *robustness* to the test set.

OoD performance of FROB for OoD few-shots from the test set but also adding a general OoD dataset. Table 4 shows the OoD detection performance of FROB for OoD few-shots from the test dataset, adding a

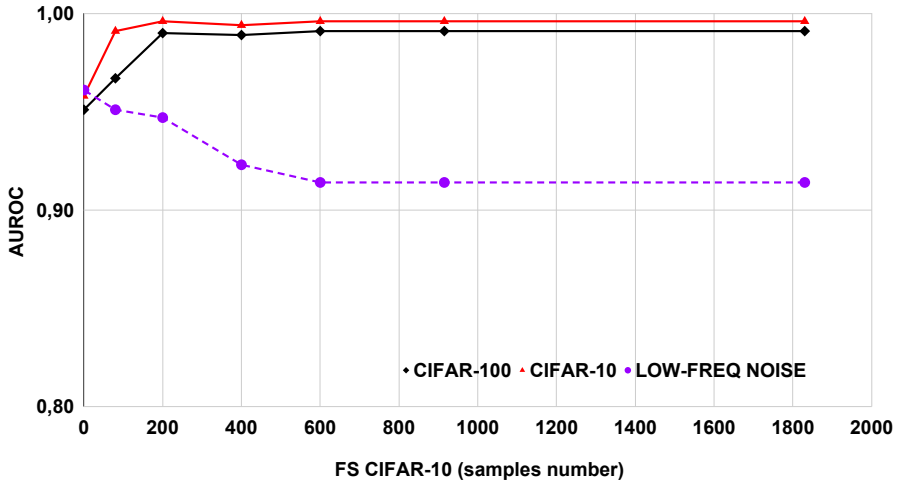

Fig. 3: Evaluation of FROB using the learned distribution boundary, $O(\mathbf{z})$, and few-shot OoD samples from CIFAR-10 in AUROC. The normal class is SVHN.

Table 4: OoD performance of FROB using the learned $O(\mathbf{z})$, OoD few-shots, and a general OoD dataset following the procedure in [18,3] resulting in 73257 samples, evaluated on SVHN. The normal class is CIFAR-10. *FS is Few-Shots.*

OoD FS	Outlier	OoD TEST	AUROC	AAUROC	GAUROC
SVHN: 1830	✓	SVHN	0.994	0.994	0.972
SVHN: 915	✓	SVHN	0.993	0.993	0.333
SVHN: 732	✓	SVHN	0.990	0.990	0.010
SVHN: 457	✓	SVHN	0.997	0.997	0.807
SVHN: 100	✓	SVHN	0.992	0.992	0.896
SVHN: 80	✓	SVHN	0.981	0.981	0.922
SVHN: 80	–	SVHN	0.995	0.995	0.928

general-purpose OoD dataset [18,3,5]. Compared to Table 2, FROB without the OoD dataset achieves *higher* AUC-metrics, and this is important. This happens because of including our proposed self-generated distribution boundary, $O(\mathbf{z})$, in our training. Adding a general-purpose OoD dataset leads to far-OoD samples which are not task-specific and might be irrelevant [17]. These far-OoD samples from the general benchmark OoD dataset are *far away* from the boundary of the

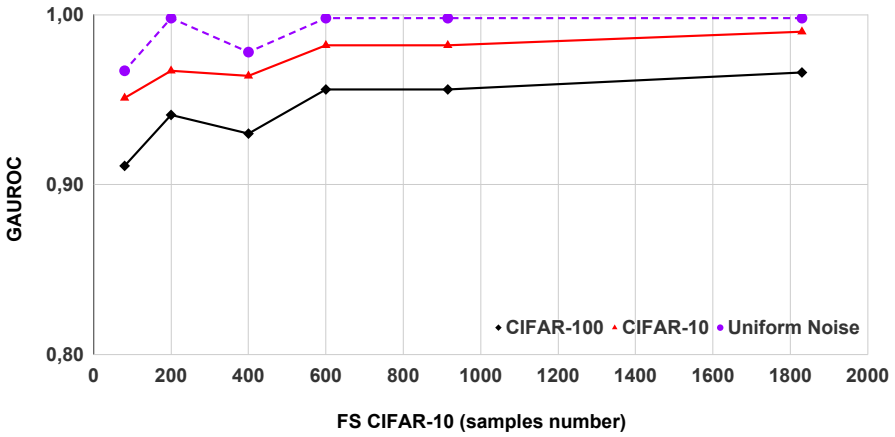

Fig. 4: FROB for normal SVHN in GAUROC with $O(\mathbf{z})$ and a variable number of OoD CIFAR-10 few-shots, tested on CIFAR-100, CIFAR-10, and Uniform Noise.

support of the normal class distribution, have high point-to-set distance as measured by the second loss term of our loss function in (2), are unevenly scattered in the data space, and are non-uniformly dispersed. Notably, in Table 2, compared to Table 4, the AUROC of FROB for normal CIFAR-10 is 0.996 and 0.995 for 100 and 80 OoD few-shots from SVHN respectively, while the AUROC of FROB with a general OoD dataset is 0.992 and 0.981 respectively. This is an important finding implying that a general OoD dataset is not needed, and that FROB with the self-generated $O(\mathbf{z})$ achieves state-of-the-art performance for few-shot OoD detection when the OoD few-shots originate from the test set. We hypothesise that the general OoD set is not required because $O(\mathbf{z})$ generates samples that are out of the data distribution that well cover the space between these samples and the normal (in-distribution) class. An external outlier OoD dataset likely provides samples that are *further out* and dispersed, and not task-specific.

We have thus shown in Table 4 that when FROB with the learned boundary, $O(\mathbf{z})$, is used during training, then the use of a general OoD dataset is not needed. Next, Fig. 4 shows the performance of FROB for normal SVHN and a variable number of OoD CIFAR-10 few-shots. In Figs. 4 and 3, compared to Fig. 2, we show that FROB achieves better performance for normal SVHN, compared to for normal CIFAR-10, in all AUC-type metrics, on the unseen CIFAR-100.

FROB compared to baselines. We compare our proposed FROB model to baselines for OoD detection. We focus on *all* the AUROC, AAUROC, and GAUROC, on the robustness of the models, and on the worst-case OoD detection performance using l_∞-norm perturbations for each of the OoD data samples.

Table 5: Performance of FROB with the self-generated $O(\mathbf{z})$, normal class C10, and general OoD set following the procedure in [18,3]. Comparison to baselines.

MODEL	$O(\mathbf{z})$	OoD DATASET	TEST	AUROC	AAUROC	GAUROC
FROB	✓	SVHN:1830	SVHN	**0.997**	**0.997**	**0.990**
FROB	✓	[18,3],SVHN:1830	SVHN	**0.994**	**0.994**	**0.972**
CCC	–	SVHN	SVHN	0.999	0.000	0.000
CEDA	–	[18,3]	SVHN	0.979	0.257	0.000
OE	–	[18,3]	SVHN	0.976	0.70	0.000
ACET	–	[18,3]	SVHN	0.966	0.880	0.000
GOOD	–	[18,3]	SVHN	0.757	0.589	0.569

We examine the OoD detection performance of the baseline models CCC, CEDA, [5], ACET, and GOOD when using C10 as the normal class, a general OoD dataset [18,3], and the SVHN OoD dataset. We evaluate these baseline models on the SVHN set. FROB outperforms baselines, specifically when the three evaluation metrics AUROC, AAUROC, and GAUROC are considered.

4.2 Ablation Studies

Removing $O(\mathbf{z})$. We remove the learned distribution boundary, $O(\mathbf{z})$, in a model we term FROBInit. We compare with FROB using OoD few-shots from SVHN, using 1830 samples, in Table 6. The OoD detection performance of FROB in AUROC in Table 6 is 0.997 and that of FROBInit, which does not use the learned boundary, $O(\mathbf{z})$, is 0.847. FROB outperforms FROBInit in all AUC-based metrics, by approximately 18% in AUROC and AAUROC and 36% in GAUROC. These results demonstrate the effectiveness and efficacy of FROB.

FROB generating the boundary, $O(\mathbf{z})$, leads to robustness to the number of OoD few-shots. Most existing methods from the literature are sensitive to the number of OoD few-shots. We demonstrate this sensitivity in Figs. 5 and 6, where we examine the performance of FROBInit which lacks the generator of boundary samples by varying the number of few-shot outliers. We also compare with FROB. Comparing Figs. 5 and 6 with Figs. 2 and 3, we see that ablating $O(\mathbf{z})$ leads to loss of robustness to a small number of few-shots.

In Figs. 5 and 6, the performance of FROBInit without the learned distribution boundary, $O(\mathbf{z})$, is not robust for few-shot OoD detection, i.e. for few-shots less than approximately 1800 samples. In Figs. 2 and 3, compared to Fig. 5, FROB achieves robust OoD detection performance as the number of OoD few-shots decreases. This indicates that $O(\mathbf{z})$ is effective and FROB is robust to the number of OoD few-shots, even to a small number of few-shot samples. We have

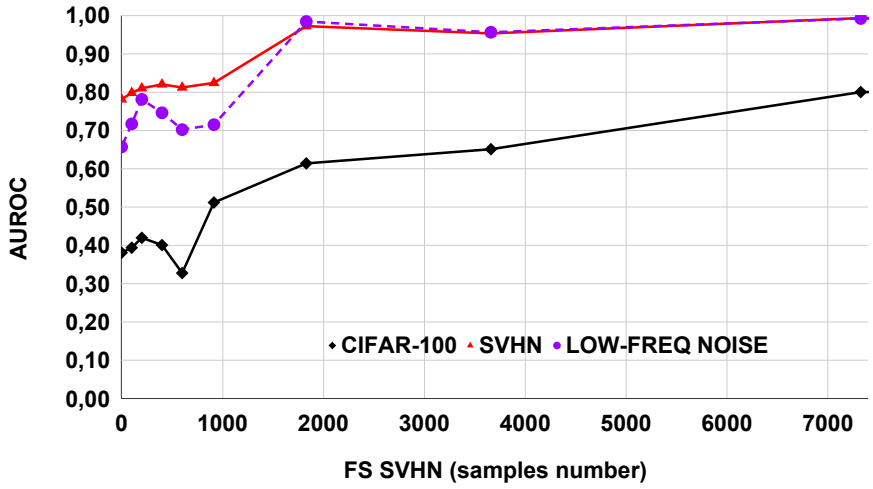

Fig. 5: OoD performance of FROBInit in AUROC, for the normal class CIFAR-10, without $O(\mathbf{z})$ and using OoD few-shots of variable number from SVHN.

Table 6: OoD performance of FROB with the learned distribution boundary, $O(\mathbf{z})$, and 1830 OoD samples from SVHN without and with a general OoD dataset following the procedure in [18,3]. The normal class is CIFAR-10.

MODEL	$O(\mathbf{z})$	OoD LOW-SHOTS	AUROC	AAUROC	GAUROC
FROB	✓	SVHN: 1830	**0.997**	**0.997**	**0.990**
FROBINIT	–	SVHN: 1830	0.847	0.847	0.728

shown that when the self-generated distribution boundary, $O(\mathbf{z})$, is not used, the OoD performance in AUROC *decreases* as the number of OoD few-shots decreases. The self-generated distribution boundary of FROB leads to a specific selection of anomalous samples that do not allow unfilled space in the data space, between the learned negatives and the normal class. FROB, because it generates samples on the distribution boundary, shows a more robust and improved OoD performance to the number of OoD few-shots when compared to FROBInit.

Further evaluation of FROBInit and its Breaking Point. To show the benefit of our proposed FROB model using our learned distribution boundary samples, $O(\mathbf{z})$, in (2), we now continue the evaluation of FROBInit in this ablation study analysis. We have demonstrated in Figs. 5 and 6 that the performance of FROBInit without the self-produced $O(\mathbf{z})$ data samples, when the normal class is the CIFAR-10 dataset, with a variable number of OoD few-shot

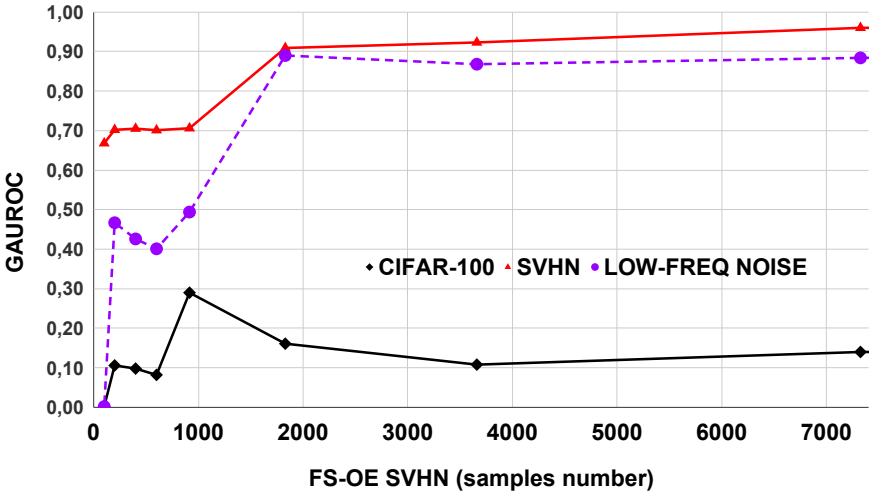

Fig. 6: Performance of FROBInit, without $O(\mathbf{z})$, in GAUROC for the normal class CIFAR-10, and with a variable number of OoD few-shots from SVHN.

samples from the SVHN dataset, when evaluated on different image datasets, decreases as the number of the few-shots of OoD data decreases.

The Break Point threshold at AUROC 0.5 is reached for approximately 800 few-shots for CIFAR-100. When the learned distribution boundary, $O(\mathbf{z})$, is not used, we do not achieve a robust performance for decreasing few-shots. The performance falls with the decreasing number of few-shots: *steep* decline for low-shots less than 1830 samples, tested on SVHN and on Low-Frequency Noise.

5 Conclusion

We have proposed FROB which uses the self-generated support boundary of the normal class distribution to improve few-shot OoD detection. FROB tackles the few-shot problem using joint classification and OoD detection. Our contribution is the combination of the generated boundary in a self-supervised learning manner and the imposition of low confidence at this learned boundary leading to improved robust few-shot OoD detection performance. To improve robustness, FROB generates strong adversarial samples on the boundary, and enforces samples from OoD and on the boundary to be less confident. By including the self-produced boundary, we reduce the threshold linked to the model's few-shot robustness. FROB redesigns, restructures, and streamlines the use of general OoD datasets to work for few-shot samples. Our proposed FROB model performs classification and few-shot OoD detection with a high level of robustness in the real world, in the wild. FROB maintains the OoD performance approximately constant, independent of the few-shot number. The performance of FROB

with the self-supervised boundary is robust and effective. Its performance is approximately stable as the OoD low- and few-shots decrease in number, while the performance of FROBInit, which is without $O(\mathbf{z})$, sharply falls as the few-shots decrease in number. The evaluation of FROB on several datasets, including the ones dissimilar to training and few-shot sets, shows that it is effective, achieves competitive state-of-the-art performance and generalization to unseen anomalies, with applicability to unknown, in the wild, test datasets, and outperforms baselines in the few-shot anomaly detection setting, in AUC-type metrics.

Acknowledgement. This work was supported by the Engineering and Physical Sciences Research Council of the UK (EPSRC) Grant EP/S000631/1 and the UK MOD University Defence Research Collaboration (UDRC) in Signal Processing.

References

1. S. Sheynin, S. Benaim, and L. Wolf. A Hierarchical Transformation-Discriminating Generative Model for Few Shot Anomaly Detection Generative models multi-class classification. In International Conference on Computer Vision (ICCV). 2021.
2. K. Wang, P. Vicol, E. Triantafillou, and R. Zemel. Few-shot Out-of-Distribution Detection. Workshop ICML, Uncertainty and Robustness in Deep Learning. 2020.
3. J. Bitterwolf, A. Meinke, and M. Hein. Certifiably adversarially robust detection of out-of-distribution data. Neural Information Processing Systems (NeurIPS). 2020.
4. M. Hein, M. Andriushchenko, and J. Bitterwolf. Why ReLU networks yield high-confidence predictions far away from the training data and how to mitigate the problem. Conference on Computer Vision and Pattern Recognition (CVPR). 2019.
5. D. Hendrycks, M. Mazeika, and T. Dietterich. Deep anomaly detection with outlier exposure. In International Conference on Learning Representations (ICLR). 2019.
6. N. Dionelis, M. Yaghoobi, and S. Tsaftaris. Boundary of Distribution Support Generator (BDSG): Sample Generation on the Boundary. In IEEE International Conference on Image Processing (ICIP). 2020.
7. N. Dionelis, M. Yaghoobi, and S. Tsaftaris. Tail of Distribution GAN (TailGAN): Generative-Adversarial- Network-Based Boundary Formation. In IEEE Sensor Signal Processing for Defence (SSPD). 2020.
8. V. Sehwag, M. Chiang, and P. Mittal. SSD: A unified framework for self-supervised outlier detection. International Conference Learning Representations (ICLR). 2021.
9. K. Lee, K. Lee, H. Lee, and J. Shin. Training confidence-calibrated classifiers for detecting out-of-distribution samples. In ICLR. 2018.
10. N. Dionelis, M. Yaghoobi, and S. Tsaftaris. Few-Shot Adaptive Detection of Objects of Concern Using Generative Models with Negative Retraining. In International Conference on Tools with Artificial Intelligence (ICTAI). 2021.
11. N. Dionelis, M. Yaghoobi, and S. Tsaftaris. OMASGAN: Out-of-Distribution Minimum Anomaly Score GAN for Sample Generation on the Boundary. arXiv:2110.15273. 2021.
12. Moon, J. and Kim, J. and Shin, Y. and Hwang, S.. Confidence-aware learning for deep neural networks. International Conference on Machine Learning (ICML). 2020.
13. O. Bastani, Y. Ioannou, L. Lampropoulos, D. Vytiniotis, A. Nori, and A. Criminisi. Measuring neural net robustness with constraints. In NeurIPS. 2016.

14. F. Croce and M. Hein. Reliable evaluation of adversarial robustness with an ensemble of diverse parameter-free attacks. In ICML. 2020.
15. M. Sensoy, L. Kaplan, F. Cerutti, and M. Saleki. Uncertainty-aware deep classifiers using generative models. In AAAI Conference on Artificial Intelligence. 2020.
16. J. Tack, S. Mo, J. Jeong, and J. Shin. CSI: Novelty detection via contrastive learning on distributionally shifted instances. In NeurIPS. 2020.
17. T. Jeong and H. Kim. OoD-MAML: Meta-Learning for Few-Shot Out-of-Distribution Detection and Classification. In Proc. NeurIPS. 2020.
18. N. Rafiee, R. Gholamipoorfard, N. Adaloglou, S. Jaxy, J. Ramakers, and M. Kollmann. Self-Supervised Anomaly Detection by Self-Distillation and Negative Sampling. arXiv:2201.06378. 2022.