A Multidirectional Meta-Learning Framework for Class-Generalizable Anomaly Detection

Anonymous Author
Anonymous Institution

Abstract

In this paper, we address the problem of class-generalizable anomaly detection, where the objective is to develop a unified model by focusing our learning on the available normal data and a small amount of anomaly data in order to detect the completely unseen anomalies, also referred to as the out-of-distribution (OOD) domain. Our aim is to learn a latent representation that eliminates class-specific spurious correlations, thereby enabling extrapolation to unseen classes or anomaly types. Adding to this challenge is the fact that the anomaly data is rare and costly to label. To achieve this, we propose a multidirectional meta-learning algorithm. At the inner level, the model aims to learn the manifold of normal data. At the outer level, the model is meta-tuned to maximize the margin between softmax confidence scores of normal and anomalous samples, treating normals as in-distribution (ID) and anomalies as out-of-distribution (OOD). By iteratively repeating this process over multiple episodes of predominantly normal and a small amount of unseen anomalies, we realize a multidirectional meta-learning framework. This two-level optimization, enhanced by multidirectional training, improves anomaly detection and enables stronger generalization to unseen anomaly classes.

Preliminary work. Under review by AISTATS 2026. Do not distribute.

1 Introduction

Anomaly detection (AD) has a wide spectrum of applications in various domains. In industrial settings, it is used to monitor machinery and detect early signs of equipment failure or faults. In network traffic and cybersecurity, anomaly detection identifies malicious activity, intrusions, or deviations in system logs that may indicate attacks or breaches. In climate science, it helps uncover unusual patterns in weather data, environmental monitoring, or long-term climate trends. In healthcare, anomaly detection supports early diagnosis by flagging irregularities in patient records, physiological signals, or medical imaging.

In anomaly detection, anomalies are inherently rare, heterogeneous, and constantly evolving, which makes them difficult to characterize exhaustively during training without additional retraining or fine-tuning. By contrast, normal data are abundant and comparatively stable across datasets, making it a more reliable foundation for representation learning. Putting emphasis on learning the normal data manifold allows capturing consistent structures and regularities that define "normality" across domains. Once this manifold is well learned, even small deviations caused by anomalies become more distinguishable, enabling the detection of novel or unseen attacks. Furthermore, fine-tuning with a limited set of anomalies can then act as a corrective step, improving sensitivity without requiring exhaustive coverage of all possible anomaly types.

This makes class-generalizable anomaly detection a particularly challenging and important problem to address. Adding to this challenge is the fact that normal classes vary significantly between datasets. In some cases, feature correlation of anomaly datasets lies very close to the normal data making them very difficult to separate. As a result, normal data from previously unseen domains

can often be misclassified as anomalies, leading to high false positive rates. Our goal in this paper is to design a class-generalizable few-shot and zero-shot anomaly detection algorithm that learns robust latent features from multiple classes and can generalize to entirely unseen or out-of-distribution (OOD) domains. To achieve this, we build on the well-established ODIN principle [39], which observes that in-distribution (ID) samples typically exhibit a larger softmax confidence score gradient than most OOD samples. Intuitively, even when ID and OOD samples produce similar softmax confidence scores, the ID samples tend to yield much larger gradient norms of the softmax score after input processing. This, in turn, leads to a sharper increase in the softmax score for ID data compared to OOD data. As a result, input preprocessing based on ODIN enhances the separability between ID and OOD samples, improving the reliability of anomaly detection.

The idea of meta-learning involves training in episodes that simulate deployment: in each episode, we (a) meta-train on some source domains and (b) meta-test on a held-out domain from the same pool, updating the model so it does well on both. Repeating this biases the representation toward domain-invariant cues that transfer to unseen domains at test time. Because every update is judged by performance on a different domain within the episode, the network learns features that survive domain swaps (i.e., are stable across shifts) and de-emphasizes spurious, domain-specific correlations. MAML [40] learns a model initialization that can adapt to a new task with just a few gradient steps. It optimizes a meta-objective across many tasks so that the post-adaptation loss is low. For OOD domain generalization, we form tasks from different domains and meta-train so that the model adapts quickly to a held-out domain.

Multidirection meta-learning is a concept within the field of meta-learning that involves transferring knowledge in multiple directions to improve a model's performance on new tasks. This approach is particularly effective for few-shot learning, where models must adapt to new tasks using only a small number of examples. Multidirection meta-learning distinguishes itself by not just transferring knowledge in a single direction (e.g., from a pre-trained model to a new task). Instead, it enables a more collaborative and dynamic exchange of information. In contrast to a one-way knowledge transfer from a fixed source to a new task, multidirection meta-learning allows different parts of the learning system to "teach" each other to achieve a common goal. We want to leverage multidirection meta-learning to improve few-shot and zero-shot learning tasks by learning the softmax confidence scores. Our goal is to train the meta-learner to learn the domain-invariant features that should be grouped together to maximize the performance.

Building on these two principles, we design a two-stage representation learning algorithm: first, it learns the normal manifold across domains in a shared latent space; then, it meta-tunes the representation using the ODIN principle to further separate in-distribution (normal) from out-of-distribution (anomalies).

In summary, our contribution can be summarized as follows:

- We address the challenge of out-of-distribution (OOD) generalization in anomaly detection (AD) by proposing a bi-level meta-learning algorithm. The inner level learns to separate normal data across domains, while the outer level meta-tunes this boundary with small numbers of anomaly samples using the ODIN principle to further distinguish normal from anomalous samples.
- We extend this framework to multiple metadirections by constructing episodes that combine normal data (inner level) with a minimal set of anomalies (outer level) from multiple domains (attack classes). This multitask objective relaxes the dependency on class-specific source and target distributions, enabling the meta-model to generalize effectively to unseen anomalies.
- Extensive experiments demonstrate consistent gains across key classification metrics, including precision, recall, and AUC, when evaluated on groups of OOD and previously unseen anomaly classes.

2 Related Work

Domain generalization methods are typically grouped into four main categories: domain-invariant representation learning, meta-learning, latent dimension regularization, and metric learning. The first group aims to extract features that remain stable across domains, enabling transfer to unseen settings - for example, autoencoder-based methods [1] use multidomain training with augmentation to learn shared representations,

MMD-AAE [24] aligns heterogeneous distributions through adversarial learning, while approaches like domain-specific masking [14], noise-enhanced autoencoders [15] and moment-alignment techniques [29, 18] capture both invariant and discriminative features. The second group, metalearning, improves generalization by leveraging related tasks, such as latent space projections to mitigate domain bias [30], MAML-style methods that adapt updates in latent space [40, 31], and zero-shot learning [9] that transfers knowledge from seen to unseen classes. A third direction, based on information bottleneck and metric learning, emphasizes disentangling spurious correlations, with theoretical guarantees provided by the variance–invariance–covariance framework [3], although adversarial adaptation methods remain sensitive to distribution mismatches. Early work like [6] relied on softmax statistics for error and OOD detection, later extended by [7] through unsupervised adaptation that minimizes inter-domain discrepancies, while more recent methods such as nearest-neighbor–based detection [10] and causal invariant learning [11] provide flexible distribution-agnostic strategies for cross-domain generalization. The ODIN paper Liang *et al.* [39] shows that standard softmax confidence is unreliable for detecting OOD inputs because networks often assign high probabilities to unfamiliar data. It proposes two simple test-time tweaks—temperature scaling of the softmax and a small input perturbation that amplify the confidence gap between ID and OOD samples. With these, ID examples become more confident, while OOD ones become less confident, enabling effective OOD detection using a basic threshold on the scaled softmax score. MTL-RED [37] proposes a novel classification framework that leverages regularization techniques to guide the latent space toward retaining only the most relevant features for out-of-distribution (OOD) classification. The result is a compressed, invariant representation that effectively discards *spurious* domain-specific information. The paper [38] addresses class-generalizable AD by proposing ResAD, a framework that detects anomalies in unseen classes without retraining. The key idea is to model residual features instead of raw features, which reduces inter-class variation and keeps normal residuals consistent across classes. ResAD integrates a Feature Converter, Feature Constraintor, and Feature Distribution Estimator to learn stable residual distributions, making anomalies identifiable as out-of-distribution. MetaOOD [42] is a zero-shot, unsupervised framework that uses metalearning to automatically select the most suitable OOD detection model for a new dataset without requiring labels. It leverages historical performance data and language-model embeddings of datasets and models to capture task similarities, achieving superior reliability across diverse domains.

3 Problem Formulation

Let $\mathcal{C} = \{1, 2, \ldots, K\}$ denote the set of class labels ("families"). We consider a collection of tasks $\mathcal{I} = \{I_1, \ldots, I_Q\}$, where each task $I_q \subseteq \mathcal{C}$ specifies a subset of labels used to define a domain split. For a designated reference task I_1, write its complement $I_1^c = \mathcal{C} \setminus I_1$. We are given a dataset $\mathcal{D} = \{(x, y)\}$ with $x \in \mathcal{X}$ and $y \in \mathcal{C}$, drawn from mixtures whose supports align with the sets in $\mathcal{I}$.

Let $f_\theta : \mathcal{X} \to \mathcal{Z}$ be an encoder and $h_\phi : \mathcal{Z} \to \mathbb{R}^K$ a classifier producing logits $\ell(x) = h_\phi(f_\theta(x)) \in \mathbb{R}^K$ for classes $\{1, \ldots, K\}$. For a temperature $T > 0$, we define the softmax score for class i as

$$S_i(x; T) \;=\; \frac{\exp\big([\ell(x)]_i / T\big)}{\sum_{j=1}^{K} \exp\big([\ell(x)]_j / T\big)}\,. \tag{1}$$

The predicted label is $\hat{y}(x) = \arg\max_i S_i(x; T)$, and the *maximum softmax probability* (confidence) is

$$S_{\hat{y}}(x; T) \;=\; \max_i S_i(x; T). \tag{2}$$

We adopt a first-order ODIN-style input preprocessing that inflates confidence on the current prediction. For step size $\epsilon > 0$,

$$\tilde{x} = x + \epsilon\, \mathrm{sgn}\Big(\nabla_x \log \max_{k \in \{0,1\}} S_y\big(h_\phi(f_{\theta_{\mathrm{enc}}}(x))/T\big)_k\Big). \tag{3}$$

Bilevel meta-learning objective We aim to learn parameters (θ, ϕ) and confidence controls (T, ϵ) that generalize across domain splits. *The inner loop models only ID normals*, while *OOD/anomalies appear only in the outer loop.* Confidence calibration is handled by T (and the ODIN step) at the meta level.

Inner (task) objective: Normals-only one-class binary cross-entropy (BCE). Given an ID distribution $\mathcal{P}_{\mathrm{in}}$ supported on I_1^c, we adapt the encoder by maximizing confidence on normal data via a one-class binary cross-entropy (all targets fixed to 1):

$$\theta^\star(\phi) \;=\; \arg\min_\theta \; \mathbb{E}_{x \sim \mathcal{S}_{\mathrm{tid}}}\Big[\mathrm{BCE}\big(S_y(x; T, \theta, \phi),\, 1\big)\Big]. \tag{4}$$

We fix T=1 in the inner step for stability; T is meta tuned in the outer loop.

Outer (meta) objective: In this stage, we employ small number of anomaly(OOD) samples to meta-tune the model to distinguish normal(ID) and anomaly(OOD) samples. Let $\mathcal{P}_{\text{in}}$ denote the ID-normal distribution and $\mathcal{P}_{\text{out}}$ an anomaly(OOD) distribution. Holding $\theta = \theta^\star(\phi)$ fixed, we minimize a calibrated ODIN-BCE over (ϕ, T, ϵ) to widen the ID–OOD confidence gap or otherwise we encourage high confidence on ID normals and low confidence on OOD

$$\begin{aligned}\min_{\phi,\,T,\,\epsilon} \quad \mathcal{L}_{\text{meta}}(\phi, T, \epsilon; \theta^\star) = \mathcal{L}_{\text{id}}(\phi, \theta^\star, T, \epsilon) \\ + \mathcal{L}_{\text{ood}}(\phi, \theta^\star, T, \epsilon) \\ + \alpha\, \mathcal{L}_{\text{margin}}(\phi, \theta^\star, T, \epsilon) \\ \text{s.t.} \quad \theta^\star(\phi) \ \text{ defined in Eq. (4).}\end{aligned} \tag{5}$$

The BCE terms are

$$\begin{aligned}\mathcal{L}_{\text{id}}(\phi, \theta^\star, T, \epsilon) = \mathbb{E}_{x\sim\mathcal{P}_{\text{id}}}\big[\text{BCE}\left(S_y(\tilde{x}; \theta^\star, \phi, T, \epsilon), 1\right)\big], \\ \mathcal{L}_{\text{ood}}(\phi, \theta^\star, T, \epsilon) = \mathbb{E}_{x\sim\mathcal{P}_{\text{ood}}}\big[\text{BCE}\left(S_y(\tilde{x}; \theta^\star, \phi, T, \epsilon), 0\right)\big],\end{aligned} \tag{6}$$

and the margin hinge is

$$\begin{aligned}\widehat{g} = \mathbb{E}_{x\sim\mathcal{P}_{\text{id}}}\big[S_y(\tilde{x}; \theta^\star, \phi, T, \epsilon)\big] \\ - \ \mathbb{E}_{x\sim\mathcal{P}_{\text{ood}}}\big[S_y(\tilde{x}; \theta^\star, \phi, T, \epsilon)\big], \\ \mathcal{L}_{\text{margin}}(\phi, \theta^\star, T, \epsilon) = \big[m - \widehat{g}\big]_+ .\end{aligned} \tag{7}$$

where $[a]_+ = \max\{0, a\}$, $m > 0$ is a *fixed* margin hyperparameter (ID–OOD logit gap), and $S_y(\cdot)$ denotes the calibrated normal and anomaly class logit (pre-sigmoid) evaluated at the ODIN-perturbed input $\tilde{x}$.

Equations (4)–(6) explicitly define a bi-level formulation that *separates the ID–OOD confidence margin* by minimizing a logistic loss on the ODIN score $s(\cdot)$ w.r.t. (ϕ, T, ϵ), while the inner solution $\theta^\star(\phi)$ (learned on normals only) fixes the normal representation manifold. This makes the intended margin widening precise and optimizable.

Episodic multi-task extension To emulate deployment across multiple domain splits, we form episodes over $I_q \in \mathcal{I}$, treating I_q^c as ID and I_q as OOD from different anomaly classes per episode. Let $\mathcal{P}_{\text{in}}^{(q)}$ and $\mathcal{P}_{\text{out}}^{(q)}$ be the corresponding episode distributions.

The overall multidirectional meta-objective aggregates episodes:

$$\min_{\phi,\,T,\,\epsilon} \ \frac{1}{Q}\sum_{q=1}^{Q} \mathcal{L}_{\text{meta}}^{(q)}(\phi, \theta, T, \epsilon). \tag{8}$$

4 Experiment Settings

In this section, we demonstrate the performance of our proposed model on benchmark cybersecurity and healthcare datasets.

4.1 Training Strategy

We train in short episodes that mimic a train–test shift: each episode selects a small set of ID (normal) families for the inner loop and a few-shot set of OOD families for the outer loop, with some completely unseen(OOD) anomalies reserved for the test phase. In the *inner loop*, we update the encoder and head using only clean ID (normal) data so the representation concentrates normal structure without being distracted by anomalies. In the *outer loop*, we form a balanced ID/OOD query set, apply an ODIN-style input perturbation, and optimize a binary objective on the normal logit augmented with a hinge margin that explicitly enforces ID–OOD separation. The temperature and perturbation scale are meta-tuned jointly with the classifier head to calibrate confidence on perturbed queries, optionally stabilized by a small ID buffer in the outer batch. To focus capacity where separation is weakest, we follow a hard-OOD curriculum that prioritizes challenging families while cycling coverage across episodes. When inner and outer objectives conflict, we apply PCGrad to resolve gradient interference and preserve features that benefit both stages. After each episode, we log per-family histograms and ROC/PR curves at three checkpoints (pre, after-inner, after-outer) to visualize how separation evolves, then proceed to the next family mix.

4.2 Handling Task Interference

To curb task interference in our bi-level OOD training, we decompose the objective into an inner term and per-family outer terms that include a calibrated-logit margin with gap (ID–OOD logit difference). We adopt PCGrad [41] to remove only harmful cross-task components: for any pair (i, j) with $g_i^\top g_j < 0$, we project

$$g_i \leftarrow g_i - \frac{g_i^\top g_j}{\|g_j\|_2^2}\, g_j,$$

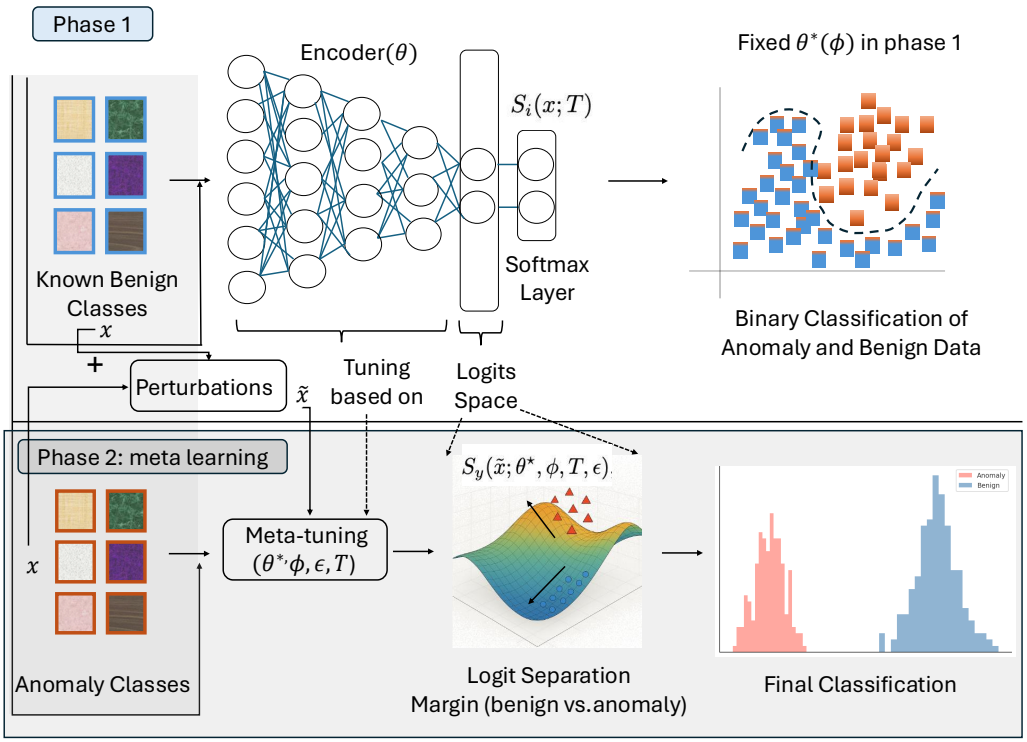


Figure 1: The Multi-directional Meta-learning Framework

and keep g_i unchanged otherwise. We apply this across (i) levels g_{in} vs. each g_k and (ii) outer families g_k vs. $g_{k'}$, using a random permutation per step to avoid bias. This preserves alignment (positive transfer) while nulling conflicting directions (negative transfer), stabilizing the margins and improving per-family AUC without ad-hoc loss reweighting.

4.3 Calibration and Evaluation

In each outer episode, we *prioritize* 2–3 *hard OOD families* from a fixed priority set $\mathcal{H} \subseteq \mathcal{C}$ (e.g., $\mathcal{H} = \{\texttt{Infil1}, \texttt{DoS1}, \texttt{Web2}\}$), selected in round-robin order and then filled occasionally from the remaining eligible families not used by the inner task. Let $K \in \{2, 3\}$ be the target number of OOD families; we select $\mathcal{O}_q \subseteq I_q$ with $|\mathcal{O}_q| = K$ when available and draw balanced OOD positives per family. This concentrates the outer signal on consistently hard anomalies while maintaining diversity across episodes.

After training, a single pooled decision threshold $\tau^\star \in [0, 1]$ is chosen on a validation pool by maximizing the F1 score:

$$\tau^\star \in \arg\max_{\tau\in[0,1]} \mathrm{F1}\big(\mathbf{1}\{p(x;T,\theta_q^\star,\phi) \geq \tau\},\, y \in I_q^c\big)\,. \tag{9}$$

Validation is performed after each episode to update $\tau^\star$ robustly, while test metrics are computed once at the end using the final $\tau^\star$. We then report per-task or per-family metrics on held-out episodes: Precision, Recall, Accuracy, F1, Average Precision, and AUC-ROC.

4.4 Hyperparameters

The training hyper-parameters include batch-size for the inner and the outer levels of the meta-learner, number of episodes, learning rate for each stage, the margin parameter, the neural network layers and the families of the various attack(anomaly) classes used at each stage. We select higher batch sizes for inner and outer optimizations and tune the learning rate in different stages.

5 Datasets and Baselines

We perform exclusive experimentation on different cybersecurity intrusion detection datasets and in the healthcare domain.

Table 1: Performance metrics (Precision (P), Recall (R), AUC, F1) for various methods across attack datasets (balanced) in CSE-CIC-IDS— Means ± std. Each of these attack datasets are considered as a held-out (OOD) while different combinations of the other attack datasets (100–600 samples) are used in the training process.

Method	BOTNET				BRUTEFORCE				GOLDENEYE			
	P	R	AUC	F1	P	R	AUC	F1	P	R	AUC	F1
CORAL	0.706 ± 0.01	0.467 ± 0.01	0.698 ± 0.01	0.562 ± 0.01	0.598 ± 0.01	0.860 ± 0.01	0.785 ± 0.01	0.705 ± 0.02	0.545 ± 0.01	0.448 ± 0.01	0.551 ± 0.01	0.492 ± 0.01
MTAE	0.545 ± 0.01	0.448 ± 0.01	0.551 ± 0.01	0.492 ± 0.01	0.653 ± 0.01	0.860 ± 0.01	0.791 ± 0.01	0.797 ± 0.01	0.667 ± 0.01	0.386 ± 0.01	0.481 ± 0.01	0.581 ± 0.01
MTL-RED	0.653 ± 0.01	0.860 ± 0.01	0.791 ± 0.01	0.742 ± 0.01	**0.930** ± 0.01	0.724 ± 0.01	0.872 ± 0.01	0.814 ± 0.01	**0.916** ± 0.01	**0.804** ± 0.01	**0.936** ± 0.01	**0.856** ± 0.01
ODIN	0.638 ± 0.01	**0.983** ± 0.01	0.891 ± 0.02	0.774 ± 0.01	0.655 ± 0.01	0.900 ± 0.01	**0.990** ± 0.01	0.792 ± 0.01	0.839 ± 0.01	0.567 ± 0.01	0.818 ± 0.01	0.677 ± 0.01
Our Model	**0.752** ± 0.01	**0.984** ± 0.01	**0.864** ± 0.01	**0.853** ± 0.01	0.884 ± 0.01	**0.980** ± 0.01	0.976 ± 0.01	**0.930** ± 0.01	0.787 ± 0.01	0.900 ± 0.01	0.900 ± 0.01	0.881 ± 0.01

Method	DDOS-HOIC				SLOWLORIS				INFILTRATION			
	P	R	AUC	F1	P	R	AUC	F1	P	R	AUC	F1
CORAL	0.523 ± 0.01	0.843 ± 0.01	0.839 ± 0.01	0.645 ± 0.01	**0.971** ± 0.01	0.733 ± 0.01	**0.981** ± 0.01	0.835 ± 0.01	0.556 ± 0.01	0.569 ± 0.01	0.584 ± 0.01	0.562 ± 0.01
MTAE	0.497 ± 0.01	0.375 ± 0.01	0.505 ± 0.01	0.428 ± 0.01	0.783 ± 0.01	0.844 ± 0.01	0.841 ± 0.01	0.875 ± 0.01	0.688 ± 0.01	0.386 ± 0.01	0.546 ± 0.01	0.494 ± 0.01
MTL-RED	0.758 ± 0.01	0.534 ± 0.01	0.604 ± 0.01	0.627 ± 0.01	0.930 ± 0.01	0.773 ± 0.01	0.929 ± 0.01	0.844 ± 0.01	**0.828** ± 0.01	0.584 ± 0.01	**0.757** ± 0.01	0.684 ± 0.01
ODIN	0.787 ± 0.01	**0.990** ± 0.01	**0.990** ± 0.01	0.881 ± 0.01	0.690 ± 0.01	**0.990** ± 0.01	0.823 ± 0.01	**0.990** ± 0.01	0.571 ± 0.01	0.571 ± 0.01	0.684 ± 0.01	0.571 ± 0.01
Our Model	**0.862** ± 0.01	**0.990** ± 0.01	0.926 ± 0.01	**0.990** ± 0.01	0.787 ± 0.01	**0.990** ± 0.01	**0.990** ± 0.01	0.882 ± 0.01	0.718 ± 0.01	**0.983** ± 0.01	0.830 ± 0.01	**0.735** ± 0.01

5.1 Dataset

- **CSE-CIC-IDS2018 [20]** This is a publicly available cybersecurity dataset that is made available by the Canadian Cybersecurity Institute (CIC). It consists of 7 major kinds of intrusion datasets namely Bruteforce, Web, Infiltration, Botnet, DoS, DDoS along with the Benign class.
- **CICIoT 2023 [20]** This is a state-of-the-art dataset for profiling, behavioral analysis, and vulnerability testing of different IoT devices with different protocols from the network traffic, consisting of 7 major attack classes - DoS, DDoS, Reconnaissance, Web, Mirai along with a Benign class for each attack categories.
- **CICIoMT 2024 [20]** This is a benchmark dataset to enable the development and evaluation of Internet of Medical Things (IoMT) security solutions. The attacks are categorized into five classes - DDoS, DoS, Reconnaissance, Spoofing, MQTT along with a Benign class for each attack categories.
- **Arrythmia** This dataset is about atrial fibrillation (also called AFib or AF) which is a quivering or irregular heartbeat (arrhythmia) that can lead to blood clots, stroke, heart failure, and other heart-related complications. The dataset contains five classes/categories: N (Normal), S (Supraventricular ectopic beat), V (Ventricular ectopic beat), F (Fusion beat), and Q (Unknown beat).

5.2 Baselines

We consider the following recent models baselines covering zero-shot and few-shot domain generalization, multi-task learning and anomaly detection.

- **Correlation Alignment for Deep Domain Adaptation (CORAL)** [34] This work has been employed for supervised domain adaptation, aligning source and target covariances to enhance OOD generalization.
- **Multi-task Autoencoder (MTAE)** [1] This encoder-decoder model optimizes reconstruction error across multiple domains in a supervised manner, jointly training sources and cross-domain data with label information in a two-stage process.
- **Out-Of-Distribution Detection in Neural Networks (ODIN)** [39] The ODIN paper proposes two simple test-time tweaks—temperature scaling of the softmax and a small input perturbation that amplify the confidence gap between ID and OOD samples. For ODIN, we consider fixed temparature scaling value(T) and ϵ for the perturbation from a possible range of values.
- **Improving Novel Anomaly Detection with Domain-Invariant Latent Representations (MTL-RED)** MTL-RED [37] propose a novel classification framework that leverages regularization techniques to guide the latent space toward retaining only the most relevant features for out-of-distribution (OOD) classification.
- **A Simple Framework for Class Generalizable Anomaly Detection (Res-AD)** This paper [38] models residual features instead of raw features, which reduces interclass variation and keeps normal residuals consistent across

Table 2: Performance metrics (Precision (P), Recall (R), AUC, F1) for various methods across datasets — Means ± std. Each of these attack datasets (balanced) in CIC-IOT/IOMT are considered as a held-out (OOD) while different combinations of the other attack datasets (100–500 samples) are used in the training process.

Method	DDoS				DoS				MIRAI			
	P	R	AUC	F1	P	R	AUC	F1	P	R	AUC	F1
CORAL	**0.986** ± 0.01	0.990 ± 0.01	0.980 ± 0.01	**0.993** ± 0.01	0.978 ± 0.01	0.875 ± 0.01	0.961 ± 0.01	0.924 ± 0.01	0.819 ± 0.01	**0.962** ± 0.01	0.863 ± 0.01	0.885 ± 0.01
MTAE	0.879 ± 0.01	0.990 ± 0.01	0.935 ± 0.01	0.964 ± 0.01	0.962 ± 0.01	0.990 ± 0.01	**0.999** ± 0.01	0.982 ± 0.01	0.880 ± 0.01	0.653 ± 0.01	0.946 ± 0.01	0.750 ± 0.01
MTL-RED	0.956 ± 0.01	0.990 ± 0.01	0.990 ± 0.01	0.977 ± 0.01	0.974 ± 0.01	**0.999** ± 0.01	0.990 ± 0.01	0.987 ± 0.01	**0.913** ± 0.01	0.866 ± 0.01	**0.959** ± 0.01	0.889 ± 0.01
ODIN	0.978 ± 0.01	0.833 ± 0.01	0.967 ± 0.01	0.900 ± 0.01	0.913 ± 0.01	0.866 ± 0.01	0.950 ± 0.01	0.889 ± 0.01	0.880 ± 0.01	0.653 ± 0.01	0.947 ± 0.01	0.780 ± 0.01
Our Model	0.974 ± 0.01	**0.990** ± 0.01	**0.990** ± 0.01	0.987 ± 0.01	**0.982** ± 0.01	0.990 ± 0.01	0.990 ± 0.01	**0.990** ± 0.01	**0.913** ± 0.01	0.866 ± 0.01	0.958 ± 0.01	**0.889** ± 0.01

Method	SPOOFING				RECONNAISSANCE				WEB			
	P	R	AUC	F1	P	R	AUC	F1	P	R	AUC	F1
CORAL	0.519 ± 0.01	**0.776** ± 0.01	0.553 ± 0.01	0.622 ± 0.01	**0.989** ± 0.01	0.615 ± 0.01	0.912 ± 0.01	0.758 ± 0.01	0.741 ± 0.01	0.619 ± 0.01	0.666 ± 0.01	0.675 ± 0.01
MTAE	0.527 ± 0.01	0.788 ± 0.01	0.742 ± 0.01	0.632 ± 0.01	0.913 ± 0.01	0.603 ± 0.01	0.709 ± 0.01	0.726 ± 0.01	0.760 ± 0.01	0.628 ± 0.01	0.751 ± 0.01	0.688 ± 0.01
MTL-RED	0.913 ± 0.02	0.603 ± 0.01	0.709 ± 0.01	0.726 ± 0.01	0.890 ± 0.01	**0.989** ± 0.01	**0.988** ± 0.01	**0.937** ± 0.01	0.824 ± 0.01	0.692 ± 0.01	0.883 ± 0.01	0.752 ± 0.01
ODIN	0.768 ± 0.01	0.414 ± 0.01	0.773 ± 0.01	0.538 ± 0.01	0.909 ± 0.01	0.822 ± 0.01	0.890 ± 0.01	0.864 ± 0.01	0.901 ± 0.01	0.603 ± 0.01	0.875 ± 0.01	0.723 ± 0.01
Our Model	**0.890** ± 0.02	0.716 ± 0.01	**0.891** ± 0.02	**0.794** ± 0.01	0.988 ± 0.01	0.825 ± 0.01	0.947 ± 0.01	0.899 ± 0.01	**0.867** ± 0.01	**0.860** ± 0.01	**0.945** ± 0.02	**0.863** ± 0.01

classes. A detailed comparative study of our work and this recent work in the context of anomaly detection is presented in our supplementary material.

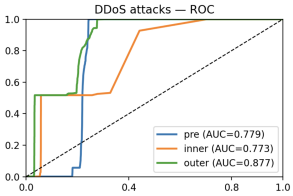
6 Ablation Study

Our ablation study mainly studies the components of our bilevel OOD learner: (i) calibrated-logit margin shaping, (ii) ODIN parameter choices (learned temperature vs fixed perturbation), and (iii) staging (pre vs. inner vs. outer). The margin penalty consistently sharpens histogram separation and boosts AP, especially when upweighting pairs with small margin. Ablating the margin increases ID and OOD overlap and raises threshold variance across families.

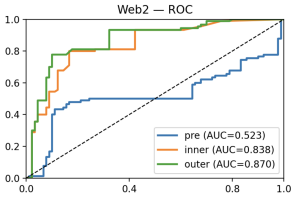
For ODIN, learning temperature T with a mild regularizer $\lambda_T \| \log T \|_2^2$ improves calibration and ROC, while attempting to learn the perturbation magnitude ϵ (in $\delta x = -\epsilon \operatorname{sign}(\nabla_x \mathcal{L}_T)$) learnable inside the gradient path introduces autograd brittleness and nonstationary steps. Fixing $\epsilon \approx 2 \times 10^{-3}$ while learning T yields the most reliable gains and avoids in-place gradient errors. Staging effects show that inner-only training helps families whose gradients align with the pretrained representation (positive transfer), and adding the outer loss increases the margin on difficult families with small gaps. An evaluation probe yields higher micro AUC on held-out splits, indicating generalization to OOD classes under representation drift and held-out validation remains the primary metric. Our best and most stable configuration combines PCGrad (across tasks and inner vs. outer), the margin penalty with $m = 0.5, \alpha = 0.1$, learned T with fixed ϵ, and a lightly unfrozen encoder, jointly maximizing macro AUC and improving per-class PR curves. While we explored learning both T and ϵ, the latter introduces additional optimization instability and engineering complexity; in practice, fixing ϵ often yields more reliable training. We evaluate across two benchmarks by rotating each dataset as held-out (OOD) while using only a small few-shot subset of anomalies from the remaining datasets (100-500 samples each class) for the upper-level meta-tuning; the encoder is first trained on normal traffic and then frozen for the inner stage. As summarized in Tables 1 and 2, our bilevel procedure improves classification metrics consistently under these cross-dataset shifts. To visualize how separation emerges, Fig. 2 reports AUC–ROC at three checkpoints—(i) a pre-tuning baseline, (ii) after the inner (ID-only) optimization, and (iii) after the outer ODIN-calibrated meta-update—showing steady gains from the margin-enforcing outer step. For testing, we construct class-balanced splits to ensure stable and comparable estimates across families and episodes.

7 Conclusion

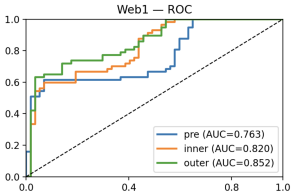
We presented a bilevel training framework for zero-shot anomaly detection that combines an inner supervised objective for fast representation shaping with an outer ODIN-style calibration objective. The outer stage learns a temperature T and enforces an adaptive margin on calibrated logits. Across episodes, we showed per-class ablations that visualize how inner updates improve class separation and how outer calibration sharpens decision thresholds. Empirically, most attack families benefit from the two-stage scheme, which yields higher AUC and AP relative to the episode baseline and demonstrates the usefulness of the margin-based calibration. We conduct a wide range of experiments on a number of



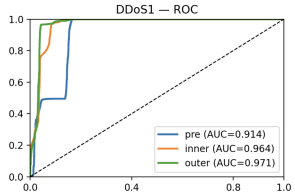
(a) AUC-ROC (DoS)



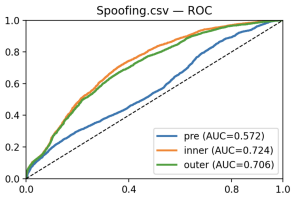
(b) AUC-ROC (WEB2)



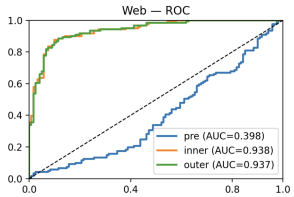
(c) AUC-ROC (WEB1)



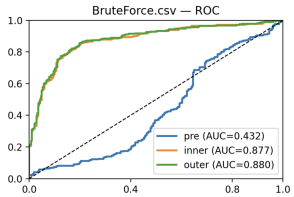
(d) AUC-ROC (DDoS1)



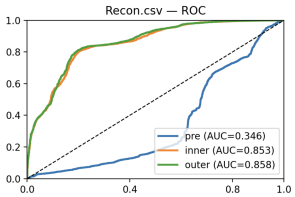
(e) AUC-ROC (Spoofing)



(f) AUC-ROC (Web)



(g) AUC-ROC (Brute-force)



(h) AUC-ROC (Reconnaissance)

Figure 2: AUC-ROC curves for different held-out (OOD) attack datasets across three benchmarks: CSE-CIC-IDS and CIC-IoMT/IOT. We plot the AUC-ROC for different held-out (OOD) attack datasets (balanced) at pre (baseline snapshot for that episode), and after inner and outer updates of the bi-level program for the held-out (OOD) classes in the datasets.

standard cybersecurity and healthcare datasets to showcase the improvements.

References

Ghifary, M., Kleijn, W.B., Zhang, M., Balduzzi, D.: Domain generalization for object recognition with multi-task autoencoders. In: Proceedings of the IEEE International Conference on Computer Vision, pp. 2551–2559 (2015)

Sathya, R., Sekar, K., Ananthi, S., Dheepa, T.: Adversarially Trained Variational Auto-Encoders With Maximum Mean Discrepancy based Regularization. In: 2022 International Conference on Knowledge Engineering and Communication Systems (ICKES), pp. 1–6. IEEE (2022)

Shwartz-Ziv, R., Balestriero, R., Kawaguchi, K., Rudner, T.G.J., LeCun, Y.: An information-theoretic perspective on variance-invariance-covariance regularization. arXiv preprint arXiv:2303.00633 (2023)

Jeon, E., Ko, W., Yoon, J.S., Suk, H.I.: Mutual information-driven subject-invariant and class-relevant deep representation learning in BCI. IEEE Transactions on Neural Networks and Learning Systems **34**(2), 739–749 (2021)

Venkataramanan, S., Psomas, B., Kijak, E., Amsaleg, L., Karantzalos, K., Avrithis, Y.: It takes two to tango: Mixup for deep metric learning. arXiv preprint arXiv:2106.04990 (2021)

Hendrycks, D., Gimpel, K.: A baseline for detecting misclassified and out-of-distribution examples in neural networks. arXiv preprint arXiv:1610.02136 (2016)

Xu, R., Chen, Z., Zuo, W., Yan, J., Lin, L.: Deep cocktail network: Multi-source unsupervised domain adaptation with category shift. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 3964–3973 (2018)

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. Advances in Neural Information Processing Systems **35**, 32598–32611 (2022)

Wang, Wei, Vincent W. Zheng, Han Yu, and Chunyan Miao. "A survey of zero-shot learning: Settings, methods, and applications." ACM Transactions on Intelligent Systems and Technology (TIST) 10, no. 2 (2019): 1-37.

Sun, Y., Ming, Y., Zhu, X., Li, Y.: Out-of-distribution detection with deep nearest neighbors. In: International Conference on Machine Learning, pp. 20827–20840. PMLR (2022)

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

Li, H., Pan, S.J., Wang, S., Kot, A.C.: Domain generalization with adversarial feature learning. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 5400–5409 (2018)

Luo, Y., Zheng, L., Guan, T., Yu, J., Yang, Y.: Taking a closer look at domain shift: Category-level adversaries for semantics consistent domain adaptation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 2507–2516 (2019)

Chattopadhyay, P., Balaji, Y., Hoffman, J.: Learning to balance specificity and invariance for in and out of domain generalization. In: ECCV 2020, pp. 301–318. Springer International Publishing (2020)

Liang, H., Zhang, Q., Dai, P., Lu, J.: Boosting the generalization capability in cross-domain few-shot learning via noise-enhanced supervised autoencoder. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 9424–9434 (2021)

Vinyals, O., Blundell, C., Lillicrap, T., Wierstra, D.: Matching networks for one shot learning. Advances in Neural Information Processing Systems **29** (2016)

Snell, J., Swersky, K., Zemel, R.: Prototypical networks for few-shot learning. Advances in Neural Information Processing Systems **30** (2017)

Lu, W., Wang, J., Li, H., Chen, Y., Xie, X.: Domain-invariant feature exploration for domain generalization. arXiv preprint arXiv:2207.12020 (2022)

Vuorio, R., Sun, S.H., Hu, H., Lim, J.J.: Multimodal model-agnostic meta-learning via task-aware modulation. Advances in Neural Information Processing Systems **32** (2019)

Canadian Institute for Cybersecurity: Public datasets for intrusion detection and anomaly detection, including CSE-CIC-IDS2018, IoT Dataset, IoMT Dataset, and Arrhythmia Dataset. Available at `https://www.unb.ca/cic/datasets/`, last accessed 2025/02/14.

Finn, C., Abbeel, P., Levine, S.: Model-agnostic meta-learning for fast adaptation of deep networks. In: International Conference on Machine Learning, pp. 1126–1135. PMLR (2017)

Guo, Y., Codella, N.C., Karlinsky, L., Codella, J.V., Smith, J.R., Saenko, K., Rosing, T., Feris, R.: A broader study of cross-domain few-shot learning. In: ECCV 2020, pp. 124–141. Springer International Publishing (2020)

Zhou, Kaiyang, Ziwei Liu, Yu Qiao, Tao Xiang, and Chen Change Loy. "Domain generalization: A survey." IEEE Transactions on Pattern Analysis and Machine Intelligence (2022).

Li, Haoliang, Sinno Jialin Pan, Shiqi Wang, and Alex C. Kot. "Domain generalization with adversarial feature learning." In Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 5400-5409. 2018.

Sung, Flood, Yongxin Yang, Li Zhang, Tao Xiang, Philip HS Torr, and Timothy M. Hospedales. "Learning to compare: Relation network for few-shot learning." In Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 1199-1208. 2018.

Yu, Shujian, Francesco Alesiani, Xi Yu, Robert Jenssen, and Jose Principe. "Measuring dependence with matrix-based entropy functional." In Proceedings of the AAAI Conference on Artificial Intelligence, vol. 35, no. 12, pp. 10781-10789. 2021.

Wang, Jindong, Cuiling Lan, Chang Liu, Yidong Ouyang, Tao Qin, Wang Lu, Yiqiang Chen, Wenjun Zeng, and Philip Yu. "Generalizing to unseen domains: A survey on domain generalization." IEEE Transactions on Knowledge and Data Engineering (2022).

Li, Ya, Xinmei Tian, Mingming Gong, Yajing Liu, Tongliang Liu, Kun Zhang, and Dacheng Tao. "Deep domain generalization via conditional invariant adversarial networks." In Proceedings of the European conference on computer vision (ECCV), pp. 624-639. 2018.

Jin, Xin, Cuiling Lan, Wenjun Zeng, and Zhibo Chen. "Feature alignment and restoration for domain generalization and adaptation." arXiv preprint arXiv:2006.12009 (2020).

Erfani, Sarah, Mahsa Baktashmotlagh, Masud Moshtaghi, Xuan Nguyen, Christopher Leckie,

James Bailey, and Rao Kotagiri. "Robust domain generalisation by enforcing distribution invariance." In Proceedings of the Twenty-Fifth International Joint Conference on Artificial Intelligence (IJCAI-16), pp. 1455-1461. AAAI Press, 2016.

Rusu, Andrei A., Dushyant Rao, Jakub Sygnowski, Oriol Vinyals, Razvan Pascanu, Simon Osindero, and Raia Hadsell. "Meta-learning with latent embedding optimization." arXiv preprint arXiv:1807.05960 (2018).

Li, Da, Yongxin Yang, Yi-Zhe Song, and Timothy Hospedales. "Learning to generalize: Meta-learning for domain generalization." In Proceedings of the AAAI conference on artificial intelligence, vol. 32, no. 1. 2018.

Seo, Seonguk, Yumin Suh, Dongwan Kim, Geeho Kim, Jongwoo Han, and Bohyung Han. "Learning to optimize domain specific normalization for domain generalization." In Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XXII 16, pp. 68-83. Springer International Publishing, 2020.

Sun, Baochen, and Kate Saenko. "Deep coral: Correlation alignment for deep domain adaptation." In Computer Vision–ECCV 2016 Workshops: Amsterdam, The Netherlands, October 8-10 and 15-16, 2016, Proceedings, Part III 14, pp. 443-450. Springer International Publishing, 2016.

Tishby, N., Pereira, F. C., Bialek, W. (2000). The information bottleneck method. arXiv preprint physics/0004057.

Kefan Dong and Tengyu Ma. First steps toward understanding the extrapolation of nonlinear models to unseen domains. arXiv preprint arXiv:2211.11719, 2022.

Roy, Padmaksha and Kevin Choi Improving Novel Anomaly Detection by Learning Domain-Invariant Representations in Latent Space ECML-PKDD 2025

Yao, X., Chen, Z., Gao, C., Zhai, G. and Zhang, C. Resad: A simple framework for class generalizable anomaly detection. Advances in Neural Information Processing Systems, 2025

Liang, S., Li, Y. and Srikant, R.,Enhancing the reliability of out-of-distribution image detection in neural networks , arXiv preprint arXiv:1706.02690

Finn, Chelsea and Abbeel, Pieter and Levine, Sergey Model-agnostic meta-learning for fast adaptation of deep networks, International conference on machine learning.

Liu, Bo and Liu, Xingchao and Jin, Xiaojie and Stone, Peter and Liu, Qiang Conflict-averse gradient descent for multi-task learning Advances in Neural Information Processing Systems, 2018

Qin, Yuehan and Zhang, Yichi and Nian, Yi and Ding, Xueying and Zhao, Yue Metaood: Automatic selection of ood detection models