Meta OOD Learning For Continuously Adaptive OOD Detection

Xinheng Wu† Jie Lu† Zhen Fang† Guangquan Zhang†
†Australian Artificial Intelligence Institute, University of Technology Sydney
Xinheng.Wu@student.uts.edu.au {Jie.Lu, Zhen.Fang, Guangquan.Zhang}@uts.edu.au *

Abstract

Out-of-distribution (OOD) detection is crucial to modern deep learning applications by identifying and alerting about the OOD samples that should not be tested or used for making predictions. Current OOD detection methods have made significant progress when in-distribution (ID) and OOD samples are drawn from static distributions. However, this can be unrealistic when applied to real-world systems which often undergo continuous variations and shifts in ID and OOD distributions over time. Therefore, for an effective application in real-world systems, the development of OOD detection methods that can adapt to these dynamic and evolving distributions is essential. In this paper, we propose a novel and more realistic setting called continuously adaptive out-of-distribution (CAOOD) detection which targets on developing an OOD detection model that enables dynamic and quick adaptation to a new arriving distribution, with insufficient ID samples during deployment time. To address CAOOD, we develop meta OOD learning (MOL) by designing a learning-to-adapt diagram such that a good initialized OOD detection model is learned during the training process. In the testing process, MOL ensures OOD detection performance over shifting distributions by quickly adapting to new distributions with a few adaptations. Extensive experiments on several OOD benchmarks endorse the effectiveness of our method in preserving both ID classification accuracy and OOD detection performance on continuously shifting distributions.

1. Introduction

Out-of-distribution (OOD) detection is vital to modern deep learning (DL) applications in the real world, especially in safety-critical applications such as autonomous vehicle control [21], and medical areas [64]. This is because DL systems may make over-confident and incorrect predictions when encountering the out-of-distribution (OOD) samples, which possess semantic labels that are distinct from those of in-distribution (ID) samples [63]. To mitigate this issue, an important problem *out-of-distribution detection* has been proposed and studied extensively [10, 17]. In OOD detection, the classifier is required to perform accurate predictions on ID samples while simultaneously identifying OOD samples. Many representative methods either utilize outputs [17, 34], feature representations [29, 45], or gradients [19, 32] to enlarge the separations between ID and OOD samples, while others [18, 39] incorporate auxiliary OOD samples to regularize the deep models during training.

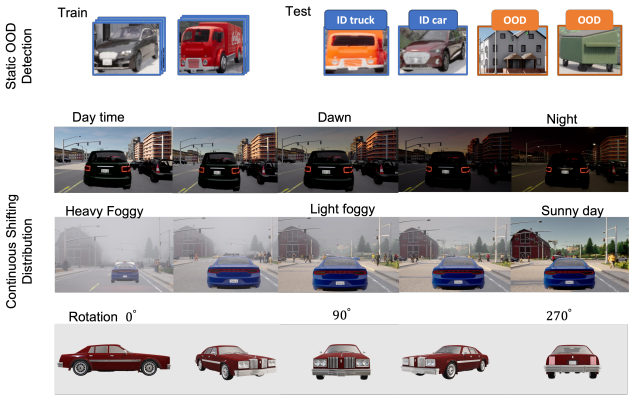

Figure 1: Difference between existing static OOD detection and a more realistic scenario of Continuously Adaptive OOD Detection (Problem 1) when test samples come from continuously shifting distribution, i.e., Definition 1.

Although OOD detection has achieved great progress, existing methods only focus on a simple scenario where ID and OOD samples are assumed to be drawn from *static* ID and OOD distributions. However, many real-world systems are changing dynamically and thus inherently exhibit continuous distribution shifts over time. Take a self-driving agent equipped with a scene recognition system as an example: in the real world, the surrounding environment (e.g., illumination, weather) may shift continuously on the road, such as from day to night, and from clean to foggy. Treating the arriving training and test samples as from *static* distributions may potentially harm the effectiveness of OOD detection, leading to the misclassification of distribution-shifted

*Corresponding to Jie Lu and Zhen Fang. The work is supported by the Australian Research Council under Discovery Grant DP200100700.

ID and OOD samples. Such scenarios highlight the limitations of current OOD detection methods when applied to continuously shifting distributions. Figure 1 highlights the difference between OOD detection on *static* distribution and when applying to continuously shifting distributions.

To tackle the continuously shifting scenarios in OOD detection, we propose a novel and realistic setting termed *continuously adaptive out-of-distribution (CAOOD) detection*, which targets on developing OOD detection method to 1) quickly adapt to the continuously shifting ID distributions and 2) detect the continuously shifting OOD samples over time. Note that during the deployment time, the ID samples may be insufficient during the adaptation process. Therefore, how to enable *dynamic* and *quick adaptation* with *insufficient* ID samples to achieve good OOD detection performance over time is core challenge of CAOOD detection.

Inspired by domain adaptation [14, 6] and meta-learning [13], we develop a novel and effective CAOOD detection method called *meta out-of-distribution learning* (MOL) to address the challenge in CAOOD detection. To dynamically and quickly adapt to the continuously shifting ID samples, we leverage the learning-to-learn paradigm [52] of meta-learning which aims to learn an internal representation for quick adaptation to a new task. Specifically, by formulating adaptations to the continuous shifting ID samples as a variety of inner tasks, we design a meta-training procedure for learning to adapt explicitly. Further, to facilitate quick adaptation with insufficient ID samples, we propose to learn a meta-representation during training, which allows us to only update the light-weighted classifier during testing while keeping the meta-representation fixed. Additionally, considering OOD samples are unavailable in training, we propose to generate OOD samples based on the shifting ID feature representations. Lastly, the classifier is trained discriminatively based on both ID and virtual OOD samples.

The contributions are summarized as follows:

- We propose a more realistic OOD detection setting called CAOOD detection to promote the applications of OOD detection techniques in real-world scenarios.
- We develop technologies based on meta-learning and domain adaptation to quickly adapt to the continuously shifting ID distributions. With these technologies, MOL is proposed to address CAOOD detection.
- We conduct extensive experiments comparing MOL with competitive OOD detection methods using various fine-tuning/adaptation strategies on 114 CAOOD detection tasks. Experiments show that MOL achieves the best performance for both ID classification and OOD detection in continuously shifting distributions.

2. Related Work

Out-of-distribution Detection methods are mainly post-hoc based adopting different scoring functions based on either logit outputs [17, 32, 34, 46, 28, 20], feature representations [45, 29, 33, 54, 47], or gradients [32, 19, 22]. These methods enjoy good applicability when deployed as they can be easily integrated into a pre-trained model without re-training. Other works utilize contrastive learning [48, 47, 55], specifically designed loss functions [60, 38, 40]. However, these methods often require sophisticated score functions but may not necessarily outperform post-hoc methods in general as noted by [63]. Another line of OOD detection methods focus on regularizing the training of classifier using either auxiliary OOD datasets or generated fake OOD samples [18, 30, 34]. These works study sampling strategies [39, 3, 31], OOD samples generation approaches [30, 51, 8], and uncertainty regularization mechanisms [50]. In addition to above strategies, other strategies have also been explored. For example, to overcome the intrinsic inconsistency of prior metrics by aggregating both the known and unknown class performance in a single performance curve, [59] propose a novel OOD detection metric named OpenAUC as the final objective function to learn OOD detector.

Above OOD detection methods neglect inevitable distribution shifts in test data over time in real-world applications. Very few OOD detection works consider such shift: one work [62] used a large unlabeled dataset containing ID samples aiming to encourage ID semantic information modeling while being robust to covariate shift. Another work [65] considers explicitly promotes the generalization capacity of the OOD detector when being evaluated on covariate-shifted ID data. Recent work [41] discovers that environmental-related features (e.g., backgrounds) significantly worsen existing OOD detection performance. These early attempts highlight existing OOD detection methods are prone to distribution shifts and consequently inadequate for realistic scenarios, which motivates our research.

Domain Adaptation (DA) aims to adapt machine learning models to unseen and different distributions [12, 64, 67, 7]. In complementary to OOD detection, DA improves models' generalization ability to covariate shift when test data share the same label set as training [35, 4]. Most existing DA methods focus on matching discrete source and target domains leveraging domain statistics [42], distance-based loss functions [37], and adversarial training [14]. Recently, different methods are developed to perform continuous domain adaptation when the target domain shifts smoothly over time [2, 53, 56]. Continuous domain adaptation is related to our problem but it does not follow an open-world assumption where test data may contain OOD samples that do not belong to the training label set. In this paper, we address a more challenging open-world learning scenario, OOD detection under continuous shifting distributions.

Open World Recognition aims to incrementally learn new information without forgetting in an open world [24, 1, 11]. This problem concentrates on zero-shot learning [44], mitigating catastrophic forgetting [25], and incremental learning [43], where unknown samples can be progressively labeled as inputs. This is different from our focus on learning an open-world classifier that can quickly adapt to continuously shifted domains online while maintaining both ID and OOD detection performance. Additional literature on the topic of meta-learning is included in the Appendix.

3. Problem Setup and Notations

OOD Detection. Let $\mathcal{X}$ denote the feature space, $\mathcal{Y} = [C]$ [1] be the label space. We consider the ID distribution $D_{X_{\rm I}Y_{\rm I}}$ as a joint distribution defined over $\mathcal{X} \times \mathcal{Y}$, where $X_{\rm I}$ and $Y_{\rm I}$ are random variables whose outputs are from spaces $\mathcal{X}$ and $\mathcal{Y}$. Given a set of n samples drawn i.i.d. from the ID distribution called ID data, $S = \{(\mathbf{x}_j, y_j)\}_{j=1}^{n} \sim D_{X_{\rm I}Y_{\rm I}}$. A classic classification model $\mathbf{f} : \mathcal{X} \to \mathbb{R}^C$, is trained on the training set S, to predict the label of an input test data [10].

During the test time of OOD detection, the test samples contain some unknown OOD samples drawn from an OOD distribution $D_{X_{\rm O}Y_{\rm O}}$, where $X_{\rm O}$ is a random variable from $\mathcal{X}$, but $Y_{\rm O}$ is a random variable whose outputs do not belong to $\mathcal{Y}$, i.e., $Y_{\rm O} \notin \mathcal{Y}$. The classical OOD detection methods aim to design an effective score function $s(\cdot; \mathbf{f}) : \mathcal{X} \to \mathbb{R}$ [32, 29, 34] and train a corresponding model $\mathbf{f}$ by ID samples S such that the following OOD detector

$$G_\gamma(\mathbf{x}; s, \mathbf{f}) = \begin{cases} \text{ID} & \text{if } s(\mathbf{x}; \mathbf{f}) \geq \gamma \\ \text{OOD} & \text{if } s(\mathbf{x}; \mathbf{f}) < \gamma \end{cases} \tag{1}$$

where γ is a threshold can distinguish ID and OOD samples accurately. In this paper, we select γ when 95% ID data is correctly classified [63, 58, 57], and use energy score [34] as the scoring function to design our OOD detector, i.e.,

$$s(\mathbf{x}; \mathbf{f}) = \log \sum_{l=1}^{C} \exp(f_l(\mathbf{x})) \tag{2}$$

where $f_l(\mathbf{x})$ is the l-th coordinate of $\mathbf{f}(\mathbf{x})$.

Continuously Adaptive OOD Detection. To tackle more realistic scenarios that the ID and OOD samples are from continuously shifting distributions over a discrete time period $T = \{t_1, t_2, ..., t_N\}$, which satisfies that $0 < t_1 < t_2 <, ..., < t_N$. We also set $T_k = \{t_1, t_2, ..., t_k\}$ and set $T_K^- = \{t_K, t_{K+1}, ..., t_N\}$, where $1 \leq k < K \leq N$. It is clear that $T_k \cup T_K^- \subset T$ and $T_k \cap T_K^- = \emptyset$. Next, the definition of continuously shifting ID and OOD distributions is given in Definition 1.

[1] We use $[N]$ to represent set $\{1, ..., N\}$. Therefore, $[C] = \{1, ..., C\}$.

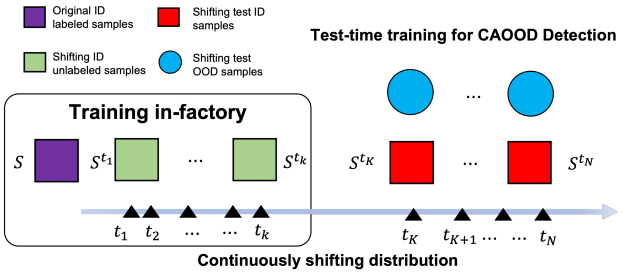

Figure 2: A heuristic illustration CAOOD problem. The target is to train an OOD detection model g_0 using S, $\{S^t\}_{t \in T_k}$ so it can obtain good OOD detection performance on unseen shifting samples $\{S^t\}_{t \in T_K^-}$ by quick adaptation.

Definition 1 (Continuously Shifting Distributions.) *Let $D^t_{X_{\rm I}Y_{\rm I}}$ and $D^t_{X_{\rm O}Y_{\rm O}}$ are ID and OOD joint distributions at time $t \in T$. We say $D^t_{X_{\rm I}Y_{\rm I}}, D^t_{X_{\rm O}Y_{\rm O}}$ are continuously shifting ID and OOD distributions, if for any $i \in [N-1]$,*

$$d(D^{t_i}_{X_{\rm I}Y_{\rm I}}, D^{t_{i+1}}_{X_{\rm I}Y_{\rm I}}) < \epsilon,$$
$$d(D^{t_i}_{X_{\rm O}}, D^{t_{i+1}}_{X_{\rm O}}) < \epsilon,$$

where $d(\cdot, \cdot)$ is a distribution metric and ϵ is a small value.

In Definition 1, the small value ϵ is used to estimate the gradual variations and quantify the continuity of continuously shifting ID and OOD distributions. We note here that distribution shift over time can be significant as it gradually accumulates. Next, we give the definition of continuously adaptive OOD detection.

Problem 1 (Continuously Adaptive OOD Detection.) *Let $D^t_{X_{\rm I}Y_{\rm I}}$ and $D^t_{X_{\rm O}Y_{\rm O}}$ be the continuously shifting ID and OOD distributions over time period T and let $D_{X_{\rm I}Y_{\rm I}}$ be the original ID distribution. Given sets of samples*

$$S = \{(\mathbf{x}_1, y_1), ..., (\mathbf{x}_n, y_n)\} \sim D_{X_{\rm I}Y_{\rm I}}, \ \ i.i.d.$$
$$S^t = \{\mathbf{x}^t_1, ..., \mathbf{x}^t_{n^t}\} \sim D^t_{X_1}, \ \ i.i.d., \text{ for each } t \in T$$

with $n^t \ll n$ if $t \in T_K^-$, the aim of continuously adaptive OOD detection is to train an initial OOD detection model g_0 using S, $\{S^t\}_{t \in T_k}$, and to quickly update the model g_{t_i} (at each time $t_i \in T_K^- = \{t_K, t_{K+1}, ..., t_N\}$) based on the previous model $g_{t_{i-1}}$ (g_0 is the initial model, i.e., $g_{t_{K-1}} = g_0$) using S, S^{t_i}. Such that for any test sample $\mathbf{x} \in \{S^t\}_{t \in T_K^-}$: 1) if $\mathbf{x}$ is from ID distribution $D^t_{X_1}$, g_{t_N} classifies $\mathbf{x}$ into correct ID label; and 2) if $\mathbf{x}$ is from OOD distribution $D^t_{X_{\rm O}}$, g_{t_N} detects $\mathbf{x}$ as OOD.

A heuristic illustration of the CAOOD problem is provided in Figure 2. The key challenge of Problem 1 is to quickly adapt to new ID distribution $D^{t_i}_{X_1}$ using insufficient ID samples S^{t_i} at time t_i. To achieve satisfied performance

across time period T_K^-, a good initial OOD detection model g_0 is indispensable for effective adaptation. In Section 4, we illustrate the proposed MOL method and detail how to obtain the initial model g_0 during the training procedure.

Notations. To facilitate better comprehension, we introduce some necessary notations. To learn a good initial OOD detection model g_0, we mainly use meta-learning techniques. Following the meta-training procedure [13], we need to sample a support set $\mathbf{S}_{\text{spt}} = \{S^t_{\text{spt}}\}_{t\in T'_k}$ from $\{S^t\}_{t\in T_k}$, where $T'_k \subset T_k$ and $S^t_{\text{spt}} \subset S^t$. We also need to sample a query set $\mathbf{S}_{\text{qry}} = \{S^t_{\text{qry}}\}_{t\in T'_k}$ from $\{S^t\}_{t\in T_k}$, where $S^t_{\text{qry}} \subset S^t$. Further, we set $\mathbf{f}_\Theta$ to be the *feature extractor*, set $\mathbf{h}_\Phi$ to be the *adapter*, and set $\mathbf{c}_W$ to be the *classifier*, where Θ, Φ and W are the parameters. Then the classification model $\mathbf{f}$ can be expressed as a function composition:

$$\mathbf{f}_{W,\Phi,\Theta} = \mathbf{c}_W \circ \mathbf{h}_\Phi \circ \mathbf{f}_\Theta.$$

Lastly, the OOD detection model G_γ can be obtained via the score function $s(\cdot; \mathbf{f}_{W,\Phi,\Theta})$, as shown in Eqs. 1 and 2.

To adapt the shifting distributions, we mainly utilize the *maximum mean discrepancy* (MMD) [15], which are introduced in following: given samples $\{\mathbf{x}_i\}_{i=1}^n$ and $\{\mathbf{x}'_j\}_{j=1}^m$,

$$\begin{aligned} &d_k(\{\mathbf{x}_i\}_{i=1}^n, \{\mathbf{x}'_j\}_{j=1}^m) \\ =&\Big\|\frac{1}{n}\sum_{i=1}^n \phi_k(\mathbf{x}_i) - \frac{1}{m}\sum_{i=1}^m \phi_k(\mathbf{x}_j)\Big\|_k, \end{aligned}$$

where k is the kernel and ϕ_k is the kernel feature map. To further improve the adaptive performance, in this work, we use the multi-kernel MMD d_{mk}, which is shown in [36].

4. Methodology

It is important to recognize that while the number of test samples S^t ($t \in T_K^-$) may be limited during deployment, we often have access to a satisfactory number of labeled training samples S as well as unlabeled samples $\{S^t\}_{t\in T_k}$ that can be used to simulate distribution shifts that may occur during deployment. This motivates us to learn a well-designed initial OOD detection model g_0 that utilizes both labeled training samples S and unlabeled samples $\{S^t\}_{t\in T_k}$ to effectively adapt to the shifting test samples S^t ($t \in T_K^-$) as defined in Problem 1.

To develop an effective initial OOD detection model g_0, we consider learning an internal meta-representation that enables efficient adaptation to the new distributions in Problem 1. Additionally, note that OOD samples are unavailable during the training process, we utilize the virtual OOD synthesize [8] techniques to generate reliable-virtual OOD samples. The virtual OOD samples emulate real OOD samples and thus enable the classifier to be trained with an additional uncertainty regularization term that incorporates both ID and virtual OOD information.

An effective strategy of learning g_0 is to draw upon the experience of meta-learning [9, 52]. Originally proposed for addressing few-shot learning [9, 23], meta-learning involves utilizing *fast adaptation* on a small number of samples. By using meta-learning [13], we can efficiently adapt the initial model to new distributions encountered during deployment. Specifically, a meta-training process is explicitly designed for learning-to-adapt by viewing the adaptation to each time $t \in T'_k$ for OOD detection as the inner tasks, thus dynamic adaptations across all time period T'_k can be regarded as one input for the outer-loop training process. Concretely, we can train the classifier by considering both ID adaptation and OOD uncertainty regularization. In the outer loop, we update the model with respect to many inner tasks across all time periods T'_k. We detail the inner learning tasks and outer-loop meta-training process in Sections 4.1 and 4.2. The entire procedures of the proposed MOL method are shown in Algorithms 1 and 2.

4.1. Inner Learning Task for MOL

ID Adaptation. In the inner task, we firstly train the adapter $\mathbf{h}_\Phi$ and classifier $\mathbf{c}_W$ to adapt to current ID distribution at time point $t \in T'_k$ using support set $\mathbf{S}_{\text{spt}}$. Specifically, given a cross-entropy loss ℓ_{ce} and multi-kernel MMD d_{mk}, we minimize the following objective.

$$\begin{aligned} \min_{\Phi,W}\left[\mathcal{L}_{\text{ce}} + \mathcal{L}^t_{\text{d}}\right] = \min_{\Phi,W}\big[&\frac{1}{n}\sum_{i=1}^n \ell_{\text{ce}}\left(\mathbf{f}_{\Theta,\Phi,W}\left(\mathbf{x}_i\right), y_i\right) \\ &+d^2_{mk}(\{\mathbf{f}_{\Theta,\Phi,W}(\mathbf{x}_i)\}_{i=1}^n, \{\mathbf{f}_{\Theta,\Phi,W}(\mathbf{x}^t_{j,\text{spt}})\}_{j=1}^{m^t})\big], \end{aligned} \tag{3}$$

where $\{\mathbf{x}^t_{j,\text{spt}}\}_{j=1}^{m^t} = S^t_{\text{spt}}$. Then at each time $t \in T'_k$, we estimate the empirical mean and covariance of training samples S for each class $c \in \mathcal{Y}$ using $\mathbf{h}_\Phi \circ \mathbf{f}_\Theta$ [29],

$$\hat{\boldsymbol{\mu}}^t_c = \frac{1}{n_c}\sum_{i:y_i=c} \mathbf{h}_\Phi \circ \mathbf{f}_\Theta\left(\mathbf{x}_i\right), \quad \hat{\boldsymbol{\Sigma}}^t = \frac{1}{n}\sum_{c=1}^C \hat{\boldsymbol{\Sigma}}^t_c, \tag{4}$$

where n_c is the number of samples S for class c, and

$$\hat{\boldsymbol{\Sigma}}^t_c = \sum_{i:y_i=c}^n \left(\mathbf{h}_\Phi \circ \mathbf{f}_\Theta\left(\mathbf{x}_i\right) - \hat{\boldsymbol{\mu}}^t_c\right)\left(\mathbf{h}_\Phi \circ \mathbf{f}_\Theta\left(\mathbf{x}_i\right) - \hat{\boldsymbol{\mu}}^t_c\right)^\top.$$

Note that the adapter $\mathbf{h}_\Phi$ is updated by the training objective in Eq. (3) at each time $t \in T'_k$.

Virtual OOD Generation. We generate virtual OOD samples based on the updated ID features $\mathbf{h}_\Phi \circ \mathbf{f}_\Theta$ at each time $t \in T'_k$. Motivated by [8, 29], we also assume that adapted class-conditional ID distribution $D^t_{\mathbf{h}_\Phi\circ\mathbf{f}_\Theta(X_{\text{I}})|Y_{\text{I}}=c}$ is similar to a multivariate Gaussian distribution, i.e.,

$$D^t_{\mathbf{h}_\Phi\circ\mathbf{f}_\Theta(X_{\text{I}})|Y_{\text{I}}=c} \approx \mathcal{N}\left(\hat{\mu}^t_c, \hat{\boldsymbol{\Sigma}}^t\right),$$

which implies that we can sample virtual OOD samples Z_c^t at the adapted feature space $\mathcal{Z} = \mathbf{h}_\Phi \circ \mathbf{f}_\Theta(\mathcal{X}) \subset \mathbb{R}^{\tilde{d}}$ to each class c in the δ-likelihood region, i.e., Z_c^t is sampled from

$$\Big\{\mathbf{z}_c^t \mid \frac{A}{|\hat{\mathbf{\Sigma}}|^{1/2}} \exp\Big((\mathbf{z}_c^t - \hat{\boldsymbol{\mu}}_c^t)\hat{\mathbf{\Sigma}}^{-1}(\mathbf{z}_c^t - \hat{\boldsymbol{\mu}}_c^t)\Big) < \delta\Big\}, \quad (5)$$

where $A = 1/(2\pi)^{\tilde{d}/2}$ and δ is a small constant to ensure that the sampled OOD samples Z_c^t are near the estimated class boundary in the adapted feature space $\mathcal{Z}$.

Inner Uncertainty Regularization. By the above steps, at each time $t \in T_k'$, we obtained the adapting ID information and the virtual OOD samples at the adapted feature space $\mathcal{Z}$. Now we add an extra regularization term such that the classifier $\mathbf{c}_W$ can classify the adapted ID samples and virtual OOD samples in the adapted feature space $\mathcal{Z}$. Specifically, the regularization term [8] is shown in following,

$$\begin{aligned}\mathcal{L}_{\text{ood}}^t = \frac{1}{C}\sum_{c=1}^{C} \mathbb{E}_{\mathbf{z}\sim\mathcal{Z}_c^t}\left[-\log\frac{1}{1+\exp^{s(\mathbf{z};\mathbf{c}_W)}}\right] \\ + \mathbb{E}_{\mathbf{x}\sim S_{\text{spt}}^t}\left[-\log\frac{\exp^{s(\mathbf{x};\mathbf{f}_{W,\Phi,\Theta})}}{1+\exp^{s(\mathbf{x};\mathbf{f}_{W,\Phi,\Theta})}}\right],\end{aligned} \quad (6)$$

where the score function s is introduced in Eq. (2). Therefore, we obtain the optimization problem for inner tasks by combining Eq. (3) and (6), i.e., for a parameter $\lambda > 0$,

$$\min_{W,\Phi} \mathcal{L}^t = \min_{W,\Phi}\big[\mathcal{L}_{\text{ce}} + \mathcal{L}_{\text{d}}^t + \lambda\mathcal{L}_{\text{ood}}^t\big]. \quad (7)$$

Note that at the earlier training stage, we optimize Φ, W only by using Eq. (3) so that a good estimation of ID distribution can be learned. Specifically, following VOS [8], we maintained a ID class-conditional queue $|Q_y|$ for each class $y \in \mathcal{Y}$ for continuous online estimation of $\hat{\mu}_c^t$ and $\hat{\Sigma}^t$. The uncertainty regularization (Eq. (7)) is introduced in the middle of the training (i.e., at a certain starting epoch E).

4.2. Learning A Meta-representation

Outer-loop Training. In Section 4.1, the inner training task only involves updating the adapter $\mathbf{h}_\Phi$ and classifier $\mathbf{c}_W$, while keeping the feature extractor $\mathbf{f}_\Theta$ fixed. In the outer loop, we target on updating $\mathbf{f}_\Theta$ corresponding to fixed Φ, W given in the inner tasks at each time $t \in T_k'$. Firstly, considering harnessing the knowledge transfer between continuous shifting distributions during time period T_k', we minimize the distribution discrepancy across the time period T_k': let $T_k' = \{t_1', t_2', ..., t_l'\}$ with $t_1' < t_2' <, ..., < t_l'$, then

$$\mathcal{L}_{\text{qry}} = \max_{i\in[l-1]} d_{mk}^2\big(\mathbf{f}_{\Theta,\Phi,W}(S_{\text{qry}}^{t_{i+1}'}), \mathbf{f}_{\Theta,\Phi,W}(S_{\text{qry}}^{t_i'})\big), \quad (8)$$

where $\mathbf{f}_{\Theta,\Phi,W}(S_{\text{qry}}^{t_i'}) = \{\mathbf{f}_{\Theta,\Phi,W}(\mathbf{x}_{j,\text{qry}}^{t_i'})\}_{j=1}^{m^{t_i'}}$, here we set $S_{\text{qry}}^{t_i'} = \{\mathbf{x}_{j,\text{qry}}^{t_i'}\}_{j=1}^{m^{t_i'}}$. Intuitively, $\mathcal{L}_{\text{qry}}$ estimates the discrepancies of continuously shifting distributions across the time period T_k'. Then, the optimization issue in outer-loop training can be represented as follows:

$$\begin{aligned}\min_{\Theta} \mathcal{L}_{\text{meta}} = \min_{\Theta}\big[\mathcal{L}_{\text{ce}} + \mathcal{L}_{\text{qry}} \\ + \frac{1}{|T_k'|}\sum_{t\in T_k'}\big(\mathcal{L}_{\text{d}}^t + \lambda\mathcal{L}_{\text{ood}}^t\big)\big],\end{aligned} \quad (9)$$

where $\mathcal{L}_{\text{d}}^t$ and $\mathcal{L}_{\text{ood}}^t$ are computed by using the samples S_{qry}^t in the query set $\mathbf{S}_{\text{qry}}$.

In Eq. (9), the inner tasks (corresponding to $\mathcal{L}_{\text{d}}^t + \lambda\mathcal{L}_{\text{ood}}^t$) across all time period T_k' are fed as one input (corresponding to $\sum_{t\in T_k'}\big(\mathcal{L}_{\text{d}}^t + \lambda\mathcal{L}_{\text{ood}}^t\big)/|T_k'|$) in the outer-loop training. Thus, we are able to learn a good initialization of meta-representation Θ such that: when encountering a new distribution during testing, a few updating steps of Φ, W result in a good performance.

Meta-testing. In the testing process, when exposed to previously unseen shifting distributions $D_{X_\text{I}Y_\text{I}}^t$ (where $t \in T_K^-$), we only need to fine-tune the adapter $\mathbf{h}_\Phi$ and classifier $\mathbf{c}_W$ on the new test samples S^t (where $t \in T_K^-$), while

Algorithm 1 . MOL in Training Process

Input: ID training samples S and $\{S^t\}_{t\in T_k}$; learning rates α, β; randomly initialized model $\mathbf{f}_{W,\Phi,\Theta} = \mathbf{c}_W \circ \mathbf{h}_\Phi \circ \mathbf{f}_\Theta$.
while *not done* **do**
  Randomly initialized adapter $\mathbf{h}_\Phi$ and classifier $\mathbf{c}_W$;
  Randomly sample $T_k' = \{t_1', t_2', ..., t_l'\}$ from T_k and sample support and query sets $\mathbf{S}_{\text{spt}}$ and $\mathbf{S}_{\text{qry}}$ from $\{S^t\}_{t\in T_k'}$;
  for $i = 1$ **to** l **do**
    Estimate $\hat{\boldsymbol{\mu}}_c^{t_i}$ and $\hat{\mathbf{\Sigma}}^{t_i}$ by using S and Eq. (4);
    Sample virtual OOD samples $Z_c^{t_i}$ by Eq. (5);
    Compute $\mathcal{L}^{t_i}$ using S, $S_{\text{spt}}^{t_i}$ and $Z_c^{t_i}$ by Eq. (7);
    Update parameters: $(\Phi, W)_{t_{i+1}} = (\Phi, W)_{t_i} - \alpha\nabla\mathcal{L}^{t_i}$;
  end for
  Compute $\mathcal{L}_{\text{meta}}$ using S and $\mathbf{S}_{\text{qry}}$ by Eq. (9);
  Update parameters: $\Theta = \Theta - \beta\nabla\mathcal{L}_{\text{meta}}$;
end while
Output: initial model $G_\gamma\left(\mathbf{x}; s, \mathbf{f}_{\Theta,\Phi,W}\right)$ by Eq. (1) meta-representation $\mathbf{f}_\Theta$.

Algorithm 2 . MOL in Testing Process

Input: ID training samples S and $\{S^t\}_{t\in T_K^-}$; learning rates α, β; meta-representation $\mathbf{f}_\Theta$; score function introduced in Eq. (2).
while *not done* **do**
  Randomly initialized adapter $\mathbf{h}_\Phi$ and classifier $\mathbf{c}_W$;
  for $i = K$ **to** N **do**
    Estimate $\hat{\boldsymbol{\mu}}_c^{t_i}$ and $\hat{\mathbf{\Sigma}}^{t_i}$ by using S and Eq. (4);
    Sample virtual OOD samples $Z_c^{t_i}$ by Eq. (5);
    Compute $\mathcal{L}^{t_i}$ using S, S^{t_i} and $Z_c^{t_i}$ by Eq. (7);
    Update parameters: $(\Phi, W)_{t_{i+1}} = (\Phi, W)_{t_i} - \alpha\nabla\mathcal{L}^{t_i}$;
  end for
  Output: model $G_\gamma\left(\mathbf{x}; s, \mathbf{f}_{\Theta,\Phi,W}\right)$ by Eq. (1).
end while

keeping the meta-representation $\mathbf{f}_\Theta$ fixed. This strategy enables fast adaptation by updating the lightweight classifier and adapter, which is useful for real-world online applications like self-driving agents. Furthermore, by incorporating the score function in Eq. (2), we can obtain an OOD detection model g_t, while updating $\mathbf{h}_\Phi$ and $\mathbf{c}_W$ at each time $t \in T_K^-$ during the testing process.

5. Experiments

5.1. Adaptive OOD Benchmark Construction

For CAOOD evaluation, we construct 3 benchmarks (e.g., using Rotation MNIST, Cifar Corruption datasets as ID datasets) derived from commonly used static OOD benchmarks (e.g., MNIST, Cifar). Rotation MNIST and Corruption datasets are frequently used in continuous domain adaptation [53, 56]. In practical applications, rotation can replicate shifts arising from camera positions, while corruptions mimic various weather conditions (e.g., fog, snow) and movements (e.g., motion). In CAOOD, we evaluate both ID accuracy and OOD detection effectiveness. Dataset details are provided in the Appendix.

Rotation MNIST [5]. This dataset has MNIST digits with various rotations from $[0, 180°]$, and for each rotation, there are 60,000 training images. In our benchmark, we use Rotation MNIST (R-MNIST) as the ID training dataset and evaluate the OOD detection performance on Rotation NOTMNIST (R-NOTMNIST) and Rotation Cifar10bw (R-Cifar10bw). There are no semantics overlaps between the ID and OOD datasets.

In our proposed MOL protocol, we consider images with rotation $0°$ as the labeled original training samples S, and images from rotation $(0, 180°]$ as continuously shifting distributions over the whole time period $(0, T]$, i.e., $\{S^t\}_{t\in T}$. In meta-training, we use images from rotation $(0 - 60°]$ as $\{S^t\}_{t\in T_k}$, and randomly sample $\mathbf{S}_{\text{spt}}$, $\mathbf{S}_{\text{qry}}$ of length 10. During meta-testing, we update our model online for rotation $\{S_{120°}, S_{126°}, i...S_{174°}\}$, i.e., $\{S^t\}_{T_K^-}$ and for each rotation, we only use 100 training samples for adaptation.

Cifar10C. This dataset consists of 15 types of corruptions (e.g., fog, brightness, motion, noise) with each demonstrating 5 levels of severity. We use Cifar10C as the ID data and test on two near OOD datasets applying the same shifting distribution: TinyimageNetC, and Cifar100C. Following the OOD benchmark literature [63], we create TinyimageNetC from a subset of the Tinyimagenet [49] where 1207 images overlap semantic labels with Cifar10 are removed.

In our protocol, we take clean images from Cifar10 as original training samples S, and consider images with various corruptions from Cifar10C as samples come from shifting distributions. Specifically, we design the continuous shifting distributions by gradually changing the severity across all corruption types, for example:

Table 1: OOD benchmarks used in our evaluation, including CAOOD datasets, near OOD, and far OOD datasets.

ID Dataset	CAOOD	Near OOD	Far OOD
R-MNIST	R-NOTMNIST R-Cifar10bw	NOTMNIST	Cifar10bw
Cifar10C	Cifar100C TinyimagenetC	Cifar100 TinyimageNet	Textures LSUN iSUN
Cifar100C	Cifar10C TinyimagenetC	Cifar10 TinyimageNet	Textures LSUN iSUN

$$\underbrace{\cdots, \mathrm{C}_{t-1}^5}_{\text{t-1 and before}} \rightarrow \underbrace{\mathrm{C}_t^1, \mathrm{C}_t^2, \mathrm{C}_t^3, \mathrm{C}_t^4, \mathrm{C}_t^5, \mathrm{C}_t^4, \mathrm{C}_t^3, \mathrm{C}_t^2, \mathrm{C}_t^1}_{\text{corruption type t, changing gradually}} \rightarrow \underbrace{\mathrm{C}_{t+1}^1, \mathrm{C}_{t+1}^2, \cdots}_{\text{t+1 and on}}$$

In meta-training, we randomly sample $\mathbf{S}_{\text{spt}}$, $\mathbf{S}_{\text{qry}}$ of length 10 (i.e., $|T_k^{'}| = 10$), from the first 7 corruptions of 63 continuously shifting distributions, i.e., $\{S^t\}_{t\in T_k}, |T_k| = 63$. In meta-testing, we adapt our model to a trajectory of length 10 that is randomly sampled from the last 7 corruptions and evaluate on unseen test images from the last 7 corruptions (i.e., frost, fog, brightness, contrast, elastic transform, pixelate, jpeg compression). Similarly, we only had access to 100 training samples in meta-testing.

Cifar100C [16]. This dataset is a variant of Cifar10C by applying the same corruptions to Cifar100. We used Cifar100C as the ID dataset and two near OOD datasets with the same continuous shifting distributions: Cifar10C and TinyimageNetC. We re-create TinyimageNetC after 2505 images sharing the same classes with Cifar100 have been removed [63]. For the MOL protocol, we applied the same training and testing fashion as used in Cifar10C.

Standard OOD Datasets. In addition, we evaluate our method's OOD detection performance on commonly used OOD datasets from a static distribution, including Cifar10 and Cifar100 [26], TinyImageNet [49], Textures [27], LSUN-Resize, LSUN-Crop [66], and iSUN [61]. Table 1 summarizes evaluated OOD detection benchmarks.

5.2. Baselines and Metrics

We compare our method with comprehensive OOD detection baselines including Maximum Softmax Probability [17], ODIN [32], Mahalanobis distance [29], energy score [34], Gram Metrics [45], Gradnorm [19], and the recent VOS [8], KNN [47] and LogitNorm [60]. To apply the above static OOD detection baselines in CAOOD, we use three different training/adapting schemes as follows.

Direct Test. The model is trained on the labeled original training samples S and then directly test on continuously shifting $\{S^t\}_{t\in T_K^-}$ for OOD detection with no adaptation.

Simple Adaptive. During training, the model is trained on the labeled original training samples S. During testing, we finetune the classifier of the model to adapt to continuously

Table 2: Main Results on Rotation MNIST. ↑ (or ↓) indicates greater (or smaller) values are preferred. For each comparable method, we report results on Direct Test / Simple Adaptive / Domain Adaptation. Only Direct Test results are reported for Gram and KNN, only Direct Test and Simple Adaptive results are reported for VOS and LogitNorm. The bold and * represent the best and second best performance and the shadow part marks our method.

Method	Rotation NOTMNIST		Rotation Cifar10bw		Average		ID Accuracy
	AUROC ↑	FPR95 ↓	AUROC ↑	FPR95 ↓	AUROC ↑	FPR95 ↓	
MSP	76.9 / 75.9 /62.3	71.9 / 70.4 / 87.6	90.1 / 89.4 / 68.8	55.6 / 48.5 / 83.6	83.5 / 82.6 / 65.6	64.8 / 59.5 / 85.6	25.6 / 28.3 / 27.2
ODIN	63.6 /76.7 / 30.8	85.2 / 72.6/ 56.7	87.3 / 88.1 / 67.6	58.2 / 50.9 / 83.9	75.5 / 82.4 / 49.2	71.7 / 61.8 / 70.3	25.6 / 28.3 / 27.2
Mahalanobis	61.2 / 75.4 / 32.9	86.1 / 75.1/ 56.9	89.3 / 90.1 / 66.7	58.1 / 49.5 / 80.8	75.3 / 82.8 / 49.8	72.1 / 62.3 / 68.9	25.6 / 28.4 / 27.7
Energy	92.6 / 91.7 / 72.7	36.1 / 39.7 / 76.1	97.5 / 95.9 / 69.1	12.6 / 19.9 / 85.2	95.1 / 93.8 / 70.9	24.4 / 29.8 / 80.7	25.6 / 28.3 / 27.2
Gram	96.1*	**10.6**	98.9*	4.6*	97.5*	**7.6**	25.6
VOS	86.7 / 93.5	57.6 / 32.3	92.7 / 96.7	31.0 / 16.4	89.7 / 95.1	44.3 / 24.4	27.7 / 30.9
LogitNorm	84.0 / 94.0	46.2 / 29.7	97.7 / **99.0**	10.5 / **4.5**	90.0 / 96.5	28.3 / 17.1	24.7 / 31.5*
KNN	95.7	16.5	91.4	25.0	93.5	20.8	24.9
MOL	**96.5**	13.9*	98.9*	9.2	**97.7**	11.5*	**35.6**

Table 3: Main Results on Cifar10 corruption. ↑ (or ↓) indicates greater (or smaller) values are preferred. For each comparable method, we report results on Direct Test / Simple Adaptive / Domain Adaptation. Only Direct Test results are reported for Gram, Gradnorm, and KNN, only Direct Test and Simple Adaptive results are reported for VOS and LogitNorm. The bold and * represent the best performance and the shadow part marks our method.

Method	Cifar100C		TinyImagenetC		Average		ID Accuracy
	AUROC ↑	FPR95 ↓	AUROC ↑	FPR95 ↓	AUROC ↑	FPR95 ↓	
MSP	53.2 / 62.2 / 54.8	94.3 / 91.6 / 92.9	59.7 / 64.0 / 52.9	92.5 / 91.2 / 94.9	56.5 / 63.1 / 53.8	93.4 / 91.4 / 93.9	38.5 / 56.2 / 35.6
ODIN	46.1 / 61.3 / 55.2	96.5 / 91.9 / 93.4	40.3 / 62.9 /52.0	99.1 / 91.5 / 94.9	43.2 / 62.1 / 53.6	97.8 / 91.7 / 94.2	38.1 / 55.6 / 35.2
Mahalanobis	46.8 / 63.6 / 55.1	96.7 / 91.3 / 93.4	47.0 / 66.0 / 53.6	98.1 / 90.6 / 93.9	46.9 / 64.8 / 54.4	97.4 / 90.9 / 93.7	38.5 / 56.2 / 35.6
Energy	52.7 / 60.8 / 56.5	94.4 / 91.4 / 93.7	61.5 / 65.0 / 60.8	91.6 / 89.0 / 94.6	57.1 / 62.9 / 58.7	93.0 / 90.2 / 94.2	38.5 / 52.9 / 35.6
Gram	55.2	89.7	66.5	**82.5**	60.9	86.1*	38.1
Gradnorm	29.5	97.2	41.3	96.5	35.4	96.9	38.5
VOS	52.6 / 53.6	95.7 / 96.5	57.3 / 59.9	97.2 / 97.1	54.9 / 56.8	96.5 / 96.8	25.9 / 31.7
LogitNorm	56.4 / 65.3*	94.5 / 88.9*	61.9 / 67.6*	92.6 / 85.9	59.1 / 66.4*	93.6 / 87.4	43.5 / 59.1*
KNN	54.1	94.4	57.7	93.5	55.9	94.0	44.2
MOL	**69.7**	**86.3**	**71.4**	85.6*	**70.6**	**85.9**	**64.1**

shifting $\{S^t\}_{t\in T_K^-}$ dynamically and test their performance. Specifically, the classifier is updated using our proposed meta-testing strategy without virtual OOD generation and uncertainty regularization (i.e., by using Eq. 3).

Domain Adaptation. The model is trained using classic domain adaptation techniques DAN [36]. Specifically, in the training process, we train the model offline by viewing S as the source domain and all shifting training samples $\{S^t\}_{t\in T_k}$ available as the target domain. During testing, we finetune the classifier of the model to adapt to continuously shifting $\{S^t\}_{t\in T_K^-}$ dynamically and test their performance (Eq. (3)). By doing so the Domain Adaptation strategy violates the CAOOD setting during training.

Note that for Gram Metrics, Gradnorm, and KNN, only the Direct Test results are reported because their scoring functions are calculated based on feature representations across all layers, which demands a finetuning on the whole model during testing. For VOS and LogitNorm, only Direct Test and Simple Adaptive results are reported, as we found it difficult to train them with Domain Adaptation strategy.

Evaluation Metrics. For OOD detection evaluation, we report 1) the area under the receiver operating characteristic curve (AUROC), which measures the model's capacity to distinguish ID and OOD samples based on varying thresholds of the predicted confidence scores; and 2) the false positive rate (FPR95) of OOD samples when the true positive rate of ID samples is fixed at 95 %, which measures the percentage of OOD samples that are incorrectly classified as ID samples when the model's sensitivity to ID samples is high. For evaluating the ID classification, we report the commonly used accuracy.

5.3. Implementation Details

On the R-MNIST dataset, we use LeNet as the backbone for all models. As for Cifar10C and Cifar100C datasets, we use WideResNet as the backbone. Note that in our method, the last fully connected layers of LeNet and WideResNet are removed for learning meta-representation during training. We set the uncertainty regularization balancing term λ as 0.015 for R-MNIST, and 0.1 for two CifarC datasets. The

Table 4: Main Results on Cifar100 corruption. ↑ (or ↓) indicates greater (or smaller) values are preferred. For each comparable method, we report results on Direct Test / Simple Adaptive / Domain Adaptation (DA). Only Direct Test results are reported for Gram, Gradnorm, and KNN, only Direct Test and Simple Adaptive results are reported for VOS and LogitNorm. The bold and * represent the best performance and the shadow part marks our method.

Method	TinyImagenetC		Cifar10C		Average		ID Accuracy
	AUROC ↑	FPR95 ↓	AUROC ↑	FPR95 ↓	AUROC ↑	FPR95 ↓	
MSP	50.9 / 61.3 / 58.1	94.8 / 91.3 / 92.5	58.2 / **64.0** / 61.2	92.9 / 91.2 / 91.3	54.6 / 62.7* / 59.7	93.4 / 91.3 / 91.9	27.8 / 41.4 / 37.7
ODIN	49.1 / 59.7 / 55.5	95.5 / 91.4 / 93.6	41.8 / 63.5 / 60.2	98.0 / 90.8* / 91.6	45.5 / 61.6 / 57.9	96.8 / 91.1* / 92.2	27.8 / 41.4 / 37.7
Mahalanobis	48.5	97.1	40.4	99.0	44.5	98.1	27.8
Energy	50.7 / 58.0/ 61.6*	94.9 / 92.1 / 91.0*	58.7 / 63.6* / 61.0	92.9 / 91.4 / 92.6	54.7 / 60.8 / 61.3	93.9 / 91.8 / 91.8	27.8 / 41.4 / 37.7
Gram	47.1	95.9	49.6	95.6	48.4	95.8	27.8
VOS	49.9 / 52.1	96.0 / 96.3	52.6 / 54.8	97.7 / 98.0	51.3 / 53.5	96.9 / 97.2	26.1 / 43.1*
LogitNorm	53.0 / 58.3	94.1 / 92.2	58.8 / 63.1	93.2 / 91.9	55.9 / 60.7	93.7 / 92.0	23.4 / 38.5
KNN	51.7	95.1	50.0	94.9	50.9	95.0	22.9
MOL	**69.4**	**88.7**	63.1	**89.0**	**66.3**	**88.9**	**57.4**

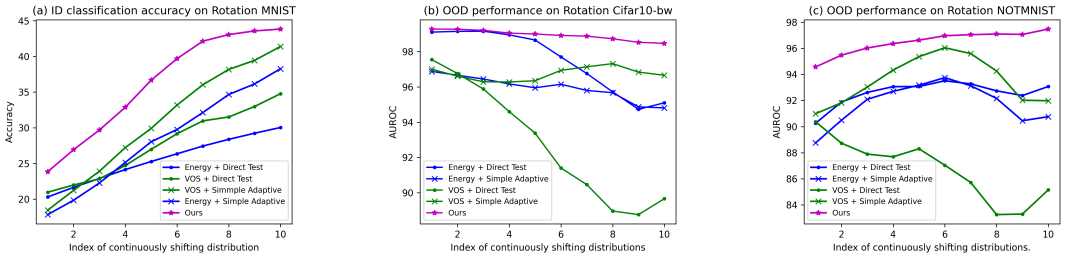

Figure 3: The ID classification (a) and OOD detection performance (b-c) on R-MNIST benchmark over continuously shifting distributions. Our method achieved the best performance on all unseen distributions, i.e., the pink line.

Table 5: Average results over 10 randomly sampled trajectories on R-MNIST.

Method	R-NOTMNIST		R-Cifar10bw		ID Accuracy
	AUROC ↑	FPR95 ↓	AUROC ↑	FPR95 ↓	
Fixed	97.9	13.9	98.9	9.2	35.6
Sampled	97.7 ±0.4	13.6 ±0.8	98.8 ±0.4	9.6 ±0.7	36.7 ±1.4

Table 6: Ablation study on R-MNIST, where w/o $\mathcal{L}_{\text{ood}}$ and w/o $\mathcal{L}_{\text{qry}}$ are variants trained without $\mathcal{L}_{\text{ood}}$ and $\mathcal{L}_{\text{qry}}$. E indicates the adding point of $\mathcal{L}_{\text{ood}}$, e.g., 1/2 indicates adding $\mathcal{L}_{\text{ood}}$ halfway through training.

Method	R-NOTMNIST		R-Cifar10bw		ID Accuracy
	AUROC ↑	FPR95 ↓	AUROC ↑	FPR95 ↓	
w/o $\mathcal{L}_{\text{ood}}$	94.1	34.9	97.6	12.2	34.2
w/o $\mathcal{L}_{\text{qry}}$	93.9	35.1	96.9	14.3	32.2
E = 1/3	99.7	7.4	99.9	5.6	28.7
E = 1/2	98.0	10.4	99.1	8.7	30.1
E = 2/3	**96.5**	**13.9**	**98.9**	**9.2**	**35.6**
E = 4/5	95.3	15.5	95.2	16.3	35.9

learning rates in inner/outer loops are set as 0.1/0.01 for R-MNIST, and 0.3/0.03 for two CifarC datasets, respectively. We use SGD with 0.9 momentum and 5×10^{-4} weight decay. On each OOD benchmark, all methods are trained using the same backbones with the same number of epochs.

5.4. Experimental Results

Results on R-MNIST. Table 2 summarizes the ID classification and OOD detection results on R-MNIST. The baseline results show that most existing static OOD detection methods cannot handle ID classification and OOD detection when test samples compass continuous distribution shifts. By using an alternate scheme when we fine-tune the model's classifier during testing (i.e., Simple Adaptive), the comparable methods improved around 2.7% in ID accuracy and reduced from 5.3% to 19.3% in FPR95 across all baselines. Such improvement endorses the dynamic adaptation of the OOD detection model when test samples come from continuously shifting distribution.

When examining the performance of the Domain Adaptation strategy, results show that ID accuracy improved over the Direct Test baselines, however, the OOD detection performance dropped heavily, from 17.9% to 26.3% across all methods in AUROC. In contrast to the Domain Adaptation strategy, our method obtained superior performance (ID accuracy: 35.6%, AUROC: 97.7%) by respecting the continuous shifting characteristic of test distributions. Note that although Gram [45] obtained the best results in FPR95, their ID accuracy (10% lower than our method) questioned the

reliability of the OOD detection performance in CAOOD.

Results on Cifar Corruption. Table 3 and 4 show the OOD detection results on Cifar10C and Cifar100C. Results show that MOL achieved the best performance on both Cifar corruption benchmarks. On Cifar10C, we outperformed Direct Test, Simple Adaptive, and Domain Adaptation strategies by 25.6%, 7.9%, and 28.5% in ID classification respectively. For OOD detection, we outperformed most comparable methods and adaptation strategies. Our AUROC outperformed the second-best result (i.e., Simple Adaptive LogitNorm) by 4.2%. On Cifar100C, we improved the 3 adaptation strategies by 29.6%, 16.0%, and 19.7% respectively in ID accuracy, while maintaining the best average OOD detection performance when compared to all other methods. Note that in our reproduction, Gram shows a catastrophic OOD performance drop and Gradnorm obtained a 100% in FPR95 when tested on the Cifar100C benchmark, which indicates the instability of their scoring functions using gradients/features across all layers when applied in CAOOD.

Performance on Continuously Shifting Distribution. Figure 3 shows the ID classification and AUROC over distributions on the Rotation MNIST benchmark. It is clear to see that the OOD detection performance of VOS and Energy decreased when the distribution continues to shift, while our method maintained a good OOD detection (i.e., the purple line). This further endorses our method's superiority in addressing the CAOOD problem by learning to quickly adapt to newly arriving test samples with shifting distributions.

Discussion on the impact of Discretization. Theoretically, modeling in continuous time intervals requires infinite samples, making it unrealistic. As such discretization (using finite samples to simulate infinite samples) is a common way to address continuous domain adaptation [56, 2]. CIDA [53] employs closed intervals which uniformly sample finite time points from these intervals and obtain training data for each sampled time point. Following CIDA, instead of adapting to one fixed discretized trajectory, in meta-training, we adapt to randomly sampled trajectories (i.e., $\mathbf{S}_{\text{spt}}$, $\mathbf{S}_{\text{qry}}$) to approximate from 0° to 60°. To further evaluate the impact of discretization on detection accuracy, we evaluate our method on 10 randomly sampled trajectories during testing. Table 5 shows the average results suggesting that the detection performance remains stable across various trajectories (fixed and sampled).

5.5. Ablation Study

A brief ablation analysis is provided in Table 6. Firstly, we build a variant of our model without $\mathcal{L}_{\text{ood}}$ by taking out the uncertainty regularization term $\mathcal{L}_{\text{ood}}$ so it reduces to focus solely on continuous ID adaptation without synthesizing virtual OOD samples for uncertainty regularization. By doing so the ID accuracy slightly dropped by 1.4% while the OOD detection performance decreases noticeably (3.8% in AUROC). Then, we conduct experiments on another variant without $\mathcal{L}_{\text{qry}}$ without the constraint term on distribution discrepancy $\mathcal{L}_{\text{qry}}$, leading to performance drops on all ID accuracy, AUROC, and FPR95. This reflects that harnessing knowledge transfer between distributions is indispensable to the CAOOD problem. Further, we studied the impact on starting point of uncertainty regularization during the whole training process. Results show that an earlier start resulted in a significant drop in ID accuracy while the OOD detection performance is good but meaningless. We suggest that a balance should be identified whenever the ID accuracy is prioritized. More ablations, experimental results, and comparisons about training/testing time refer to the Appendix.

6. Conclusion and Future Works

OOD detection has achieved great progress while facing challenges in real-world scenarios when test samples exhibit dynamic distribution shifting. Motivated by these challenges, we propose a novel and more realistic CAOOD detection and develop an effective method of MOL in addressing this problem. To the best of our knowledge, we are the first to investigate OOD detection under continuous distribution shifts. Our proposal will motivate future works to pursue new methods in addressing real-world OOD detection under continuously shifting distributions.

References

[1] Abhijit Bendale and Terrance Boult. Towards open world recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1893–1902, 2015.

[2] Andreea Bobu, Eric Tzeng, Judy Hoffman, and Trevor Darrell. Adapting to continuously shifting domains. In *International Conference on Learning Representations*, 2018.

[3] Jiefeng Chen, Yixuan Li, Xi Wu, Yingyu Liang, and Somesh Jha. Atom: Robustifying out-of-distribution detection using outlier mining. In *Machine Learning and Knowledge Discovery in Databases. Research Track: European Conference, ECML PKDD 2021, Bilbao, Spain, September 13–17, 2021, Proceedings, Part III 21*, pages 430–445. Springer, 2021.

[4] Rui Dai, Yonggang Zhang, Zhen Fang, Bo Han, and Xinmei Tian. Moderately distributional exploration for domain generalization. *International Conference on Machine Learning*, 2023.

[5] Li Deng. The mnist database of handwritten digit images for machine learning research. *IEEE Signal Processing Magazine*, 29(6):141–142, 2012.

[6] Jiahua Dong, Yang Cong, Gan Sun, Zhen Fang, and Zhengming Ding. Where and how to transfer: Knowledge aggregation-induced transferability perception for unsupervised domain adaptation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 1(2):1–18, 2021.

[7] Jiahua Dong, Yang Cong, Gan Sun, Bineng Zhong, and Xiaowei Xu. What can be transferred: Unsupervised domain adaptation for endoscopic lesions segmentation. In

IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pages 4022–4031, June 2020.

[8] Xuefeng Du, Zhaoning Wang, Mu Cai, and Yixuan Li. Vos: Learning what you don't know by virtual outlier synthesis. In *International Conference on Learning Representations*, 2022.

[9] Thomas Elsken, Benedikt Staffler, Jan Hendrik Metzen, and Frank Hutter. Meta-learning of neural architectures for few-shot learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12365–12375, 2020.

[10] Zhen Fang, Yixuan Li, Jie Lu, Jiahua Dong, Bo Han, and Feng Liu. Is out-of-distribution detection learnable? *Advances in Neural Information Processing Systems*, 35:37199–37213, 2022.

[11] Zhen Fang, Jie Lu, Anjin Liu, Feng Liu, and Guangquan Zhang. Learning bounds for open-set learning. In *International Conference on Machine Learning*, pages 3122–3132. PMLR, 2021.

[12] Zhen Fang, Jie Lu, Feng Liu, and Guangquan Zhang. Semi-supervised heterogeneous domain adaptation: Theory and algorithms. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(1):1087–1105, 2023.

[13] Chelsea Finn, Pieter Abbeel, and Sergey Levine. Model-agnostic meta-learning for fast adaptation of deep networks. In *International Conference on Machine Learning*, pages 1126–1135. PMLR, 2017.

[14] Yaroslav Ganin, Evgeniya Ustinova, Hana Ajakan, Pascal Germain, Hugo Larochelle, François Laviolette, Mario Marchand, and Victor Lempitsky. Domain-adversarial training of neural networks. *The journal of machine learning research*, 17(1):2096–2030, 2016.

[15] Arthur Gretton, Karsten M Borgwardt, Malte J Rasch, Bernhard Schölkopf, and Alexander Smola. A kernel two-sample test. *The Journal of Machine Learning Research*, 13(1):723–773, 2012.

[16] Dan Hendrycks and Thomas Dietterich. Benchmarking neural network robustness to common corruptions and perturbations. In *International Conference on Learning Representations*, 2019.

[17] Dan Hendrycks and Kevin Gimpel. A baseline for detecting misclassified and out-of-distribution examples in neural networks. In *International Conference on Learning Representations*, 2016.

[18] Dan Hendrycks, Mantas Mazeika, and Thomas Dietterich. Deep anomaly detection with outlier exposure. In *International Conference on Learning Representations*, 2018.

[19] Rui Huang, Andrew Geng, and Yixuan Li. On the importance of gradients for detecting distributional shifts in the wild. *Advances in Neural Information Processing Systems*, 34:677–689, 2021.

[20] Rui Huang and Yixuan Li. Mos: Towards scaling out-of-distribution detection for large semantic space. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8710–8719, 2021.

[21] Xiaowei Huang, Daniel Kroening, Wenjie Ruan, James Sharp, Youcheng Sun, Emese Thamo, Min Wu, and Xinping Yi. A survey of safety and trustworthiness of deep neural networks: Verification, testing, adversarial attack and defence, and interpretability. *Computer Science Review*, 37:100270, 2020.

[22] Conor Igoe, Youngseog Chung, Ian Char, and Jeff Schneider. How useful are gradients for ood detection really? *arXiv preprint arXiv:2205.10439*, 2022.

[23] Taewon Jeong and Heeyoung Kim. Ood-maml: Meta-learning for few-shot out-of-distribution detection and classification. *Advances in Neural Information Processing Systems*, 33:3907–3916, 2020.

[24] KJ Joseph, Salman Khan, Fahad Shahbaz Khan, and Vineeth N Balasubramanian. Towards open world object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5830–5840, 2021.

[25] KJ Joseph, Jathushan Rajasegaran, Salman Khan, Fahad Shahbaz Khan, and Vineeth N Balasubramanian. Incremental object detection via meta-learning. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(12):9209–9216, 2021.

[26] Alex Krizhevsky, Geoffrey Hinton, et al. Learning multiple layers of features from tiny images. 2009.

[27] Gustaf Kylberg. *Kylberg texture dataset v. 1.0*. 2011.

[28] Balaji Lakshminarayanan, Alexander Pritzel, and Charles Blundell. Simple and scalable predictive uncertainty estimation using deep ensembles. *Advances in neural information processing systems*, 30, 2017.

[29] Kimin Lee, Kibok Lee, Honglak Lee, and Jinwoo Shin. A simple unified framework for detecting out-of-distribution samples and adversarial attacks. *Advances in neural information processing systems*, 31, 2018.

[30] KIMIN LEE, Kibok Lee, Honglak Lee, and Jinwoo Shin. Training confidence-calibrated classifiers for detecting out-of-distribution samples. In *International Conference on Learning Representations*, 2018.

[31] Yi Li and Nuno Vasconcelos. Background data resampling for outlier-aware classification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13218–13227, 2020.

[32] Shiyu Liang, Yixuan Li, and R Srikant. Enhancing the reliability of out-of-distribution image detection in neural networks. In *International Conference on Learning Representations*, 2018.

[33] Ziqian Lin, Sreya Dutta Roy, and Yixuan Li. Mood: Multi-level out-of-distribution detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15313–15323, 2021.

[34] Weitang Liu, Xiaoyun Wang, John Owens, and Yixuan Li. Energy-based out-of-distribution detection. *Advances in neural information processing systems*, 33:21464–21475, 2020.

[35] Ziwei Liu, Zhongqi Miao, Xingang Pan, Xiaohang Zhan, Dahua Lin, Stella X Yu, and Boqing Gong. Open compound domain adaptation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12406–12415, 2020.

[36] Mingsheng Long, Yue Cao, Jianmin Wang, and Michael Jordan. Learning transferable features with deep adaptation networks. In *International Conference on Machine Learning*, pages 97–105. PMLR, 2015.

[37] Mingsheng Long, Han Zhu, Jianmin Wang, and Michael I Jordan. Deep transfer learning with joint adaptation networks. In *International Conference on Machine Learning*, pages 2208–2217. PMLR, 2017.

[38] Andrey Malinin and Mark Gales. Predictive uncertainty estimation via prior networks. *Advances in neural information processing systems*, 31, 2018.

[39] Yifei Ming, Ying Fan, and Yixuan Li. Poem: Out-of-distribution detection with posterior sampling. In *International Conference on Machine Learning*, pages 15650–15665. PMLR, 2022.

[40] Yifei Ming, Yiyou Sun, Ousmane Dia, and Yixuan Li. Cider: Exploiting hyperspherical embeddings for out-of-distribution detection. *arXiv preprint arXiv:2203.04450*, 2022.

[41] Yifei Ming, Hang Yin, and Yixuan Li. On the impact of spurious correlation for out-of-distribution detection. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 10051–10059, 2022.

[42] Sinno Jialin Pan, Ivor W Tsang, James T Kwok, and Qiang Yang. Domain adaptation via transfer component analysis. *IEEE Transactions on Neural Networks*, 22(2):199–210, 2010.

[43] Juan-Manuel Perez-Rua, Xiatian Zhu, Timothy M Hospedales, and Tao Xiang. Incremental few-shot object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13846–13855, 2020.

[44] Shafin Rahman, Salman H Khan, and Fatih Porikli. Zero-shot object detection: Joint recognition and localization of novel concepts. *International Journal of Computer Vision*, 128:2979–2999, 2020.

[45] Chandramouli Shama Sastry and Sageev Oore. Detecting out-of-distribution examples with gram matrices. In *International Conference on Machine Learning*, pages 8491–8501. PMLR, 2020.

[46] Yiyou Sun, Chuan Guo, and Yixuan Li. React: Out-of-distribution detection with rectified activations. *Advances in Neural Information Processing Systems*, 34:144–157, 2021.

[47] Yiyou Sun, Yifei Ming, Xiaojin Zhu, and Yixuan Li. Out-of-distribution detection with deep nearest neighbors. In *International Conference on Machine Learning*, pages 20827–20840. PMLR, 2022.

[48] Jihoon Tack, Sangwoo Mo, Jongheon Jeong, and Jinwoo Shin. Csi: Novelty detection via contrastive learning on distributionally shifted instances. *Advances in neural information processing systems*, 33:11839–11852, 2020.

[49] Antonio Torralba, Rob Fergus, and William T Freeman. 80 million tiny images: A large data set for nonparametric object and scene recognition. *IEEE Transactions on pattern analysis and machine intelligence*, 30(11):1958–1970, 2008.

[50] Joost Van Amersfoort, Lewis Smith, Yee Whye Teh, and Yarin Gal. Uncertainty estimation using a single deep deterministic neural network. In *International Conference on Machine Learning*, pages 9690–9700. PMLR, 2020.

[51] Sachin Vernekar, Ashish Gaurav, Vahdat Abdelzad, Taylor Denouden, Rick Salay, and Krzysztof Czarnecki. Out-of-distribution detection in classifiers via generation. *arXiv preprint arXiv:1910.04241*, 2019.

[52] Ricardo Vilalta and Youssef Drissi. A perspective view and survey of meta-learning. *Artificial intelligence review*, 18:77–95, 2002.

[53] Hao Wang, Hao He, and Dina Katabi. Continuously indexed domain adaptation. In *Proceedings of the 37th International Conference on Machine Learning*, pages 9898–9907, 2020.

[54] Haoqi Wang, Zhizhong Li, Litong Feng, and Wayne Zhang. Vim: Out-of-distribution with virtual-logit matching. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4921–4930, 2022.

[55] Haotao Wang, Aston Zhang, Yi Zhu, Shuai Zheng, Mu Li, Alex J Smola, and Zhangyang Wang. Partial and asymmetric contrastive learning for out-of-distribution detection in long-tailed recognition. In *International Conference on Machine Learning*, pages 23446–23458. PMLR, 2022.

[56] Qin Wang, Olga Fink, Luc Van Gool, and Dengxin Dai. Continual test-time domain adaptation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7201–7211, 2022.

[57] Qizhou Wang, Feng Liu, Yonggang Zhang, Jing Zhang, Chen Gong, Tongliang Liu, and Bo Han. Watermarking for out-of-distribution detection. In *Advances in Neural Information Processing Systems*, 2022.

[58] Qizhou Wang, Junjie Ye, Feng Liu, Quanyu Dai, Marcus Kalander, Tongliang Liu, Jianye Hao, and Bo Han. Out-of-distribution detection with implicit outlier transformation. In *International Conference on Learning Representations*. OpenReview.net, 2023.

[59] Zitai Wang, Qianqian Xu, Zhiyong Yang, Yuan He, Xiaochun Cao, and Qingming Huang. Openauc: Towards auc-oriented open-set recognition. *Advances in Neural Information Processing Systems*, 35:25033–25045, 2022.

[60] Hongxin Wei, Renchunzi Xie, Hao Cheng, Lei Feng, Bo An, and Yixuan Li. Mitigating neural network overconfidence with logit normalization. In *International Conference on Machine Learning*, pages 23631–23644. PMLR, 2022.

[61] Jianxiong Xiao, James Hays, Krista A Ehinger, Aude Oliva, and Antonio Torralba. Sun database: Large-scale scene recognition from abbey to zoo. In *2010 IEEE computer society Conference on Computer Vision and Pattern Recognition*, pages 3485–3492. IEEE, 2010.

[62] Jingkang Yang, Haoqi Wang, Litong Feng, Xiaopeng Yan, Huabin Zheng, Wayne Zhang, and Ziwei Liu. Semantically coherent out-of-distribution detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 8301–8309, 2021.

[63] Jingkang Yang, Pengyun Wang, Dejian Zou, Zitang Zhou, Kunyuan Ding, Wenxuan Peng, Haoqi Wang, Guangyao Chen, Bo Li, Yiyou Sun, et al. Openood: Benchmarking generalized out-of-distribution detection. *Advances in Neural Information Processing Systems*, 35:32598–32611, 2022.

[64] Jingkang Yang, Kaiyang Zhou, Yixuan Li, and Ziwei Liu. Generalized out-of-distribution detection: A survey. *arXiv preprint arXiv:2110.11334*, 2021.

[65] Jingkang Yang, Kaiyang Zhou, and Ziwei Liu. Full-spectrum out-of-distribution detection. *arXiv preprint arXiv:2204.05306*, 2022.

[66] Fisher Yu, Ari Seff, Yinda Zhang, Shuran Song, Thomas Funkhouser, and Jianxiong Xiao. Lsun: Construction of a large-scale image dataset using deep learning with humans in the loop. *arXiv preprint arXiv:1506.03365*, 2015.

[67] Li Zhong, Zhen Fang, Feng Liu, Bo Yuan, Guangquan Zhang, and Jie Lu. Bridging the theoretical bound and deep algorithms for open set domain adaptation. *IEEE Transactions on Neural Networks and Learning Systems*, 2021.