

---

# Learning to Shape In-distribution Feature Space for Out-of-distribution Detection

---

**Yonggang Zhang<sup>1</sup>, Jie Lu<sup>2</sup>, Bo Peng<sup>2</sup>, Zhen Fang<sup>2</sup>, Yiu-ming Cheung<sup>1\*</sup>**

<sup>1</sup>Hong Kong Baptist University

<sup>2</sup>Australian Artificial Intelligence Institute, University of Technology Sydney

## Abstract

Out-of-distribution (OOD) detection is critical for deploying machine learning models in the open world. To design scoring functions that discern OOD data from the in-distribution (ID) cases from a pre-trained discriminative model, existing methods tend to make rigorous distributional assumptions either explicitly or implicitly due to the lack of knowledge about the learned feature space in advance. The mismatch between the learned and assumed distributions motivates us to raise a fundamental yet under-explored question: *Is it possible to deterministically model the feature distribution while pre-training a discriminative model?* This paper gives an affirmative answer to this question by presenting a Distributional Representation Learning (DRL) framework for OOD detection. In particular, DRL explicitly enforces the underlying feature space to conform to a pre-defined mixture distribution, together with an online approximation of normalization constants to enable end-to-end training. Furthermore, we formulate DRL into a provably convergent Expectation-Maximization algorithm to avoid trivial solutions and rearrange the sequential sampling to guide the training consistency. Extensive evaluations across mainstream OOD detection benchmarks empirically manifest the superiority of the proposed DRL over its advanced counterparts.

## 1 Introduction

Despite the significant progress in machine learning that has facilitated a broad spectrum of classification tasks [45, 81, 42], models often operate under a *closed-world* scenario, where test data stems from the same distribution as the training data. However, real-world applications often entail scenarios in which deployed models may encounter unseen classes of samples during training, giving rise to what is known as out-of-distribution (OOD) data. These OOD instances can potentially undermine a model’s stability and, in certain cases, inflict severe damage on its performance. Accordingly, a reliable discriminative model should not only correctly classify known In-Distribution (ID) samples but also flag any OOD input as “unknown”. This directly motivates OOD detection [31, 56, 74], which makes significant differences in ensuring the safety of decision-critical applications, e.g., autonomous driving [26], medical diagnosis [85], and cyber-security [51].

Up to now, a plethora of OOD detection algorithms have been developed recently by leveraging post-hoc analysis on the pre-trained model. The seminal work [22] leverages the maximum softmax probability (MSP), also known as the softmax confidence score, for OOD detection, which is built upon the hypothesis that OOD data should trigger relatively lower softmax confidence than that of ID data. ODIN [35] extends MSP by using temperature scaling and input perturbation to amplify the ID/OOD separability. Approaches of this category, albeit intuitive, are challenged by the observation that deep neural networks are prone to produce over-confident predictions, i.e., abnormally high softmax confidences, even though the inputs are far away from the training data [50]. As a result,

---

\*Correspondence to Yiu-ming Cheung (ymc@comp.hkbu.edu.hk)

advanced methods turn to design alternative OOD scoring functions by resorting to the stored ID patterns in gradients [25], intermediate features [2, 34, 62, 36, 57, 76, 77] and logits [21, 38, 67, 78]. However, most of these methods suffer from the lack of inherent connections and theoretical understandings [48] regarding the detectability of OOD data.

Given that OOD data, by definition, inherently diverges from ID data by means of their data density distributions, some works [37, 48, 54] are motivated to focus on constructing scoring functions that can effectively replicate the behaviour of the ID density function, which opens the door to density-based OOD detection. Despite the empirical success, the power of density-based OOD detection has yet to be fully unleashed even with the helm of feature-shaping [11, 59, 60, 72, 79, 80, 84] and neuron pruning [1, 61]. This is because current methods struggle with making a strong distributional assumption of the underlying feature space either explicitly or implicitly due to the lack of knowledge about the learned feature space in advance. The urgent need to cast off the dilemma that those pre-defined distributions fail to necessarily hold in practice [62] directly promotes the following important yet under-explored question:

*Is it possible to deterministically shape the ID feature distribution while pre-training a discriminative model?*

**Methodological Contribution.** In this paper, we propose a novel learning framework called Distributional Representation Learning (DRL) that bridges the gap between network pre-training and density-based scoring strategy. At a high level, we explicitly enforce the underlying feature space to conform to a pre-defined distribution, inspired by [41]. In this way, OOD detection can be naturally approached in an assumption-free manner, hence providing stronger flexibility and generality. When implementing our idea, it could always be *non-trivial* to parameterize the ID data distribution since the computation of normalization constants tends to be costly and even intractable [19]. Towards this dilemma, instead of introducing impractical constraints to make the normalization constants input-independent or known, we propose an online approximation of normalization constants to enable end-to-end training.

**Theoretical Contribution.** Importantly, we provide a theoretical framework that formulates DRL as an Expectation-Maximization (EM) algorithm [47]. In particular, by introducing latent variables as ID classes, we obtain the Bayes-optimal inference of the posterior distribution of the latent variables given the observed data in the E-step. In the M-step, we propose to maximize the evidence lower bound (ELBO) of the log conditional likelihood with respect to all parameters so that the optimization process can not only benefit from the classification process but also avoid leading to trivial solutions. We prove that the ELBO is bounded, and the EM algorithm contributes to the convergence of ELBO. Moreover, we rearrange the sequential sampling by constraining half of each mini-batch coinciding with the previous iteration, which deals with the inconsistency issue caused by the integration of the EM algorithm into the batch-based training routine.

**Empirical Contribution.** We extensively evaluate DRL on mainstream OOD detection benchmarks and establish state-of-the-art performance compared with three families of methods: (1) pre-trained with cross-entropy, (2) pre-trained with contrastive learning, (3) pre-trained with cross-entropy and fine-tuned with training-time regularizations. For example, on CIFAR-10, DRL achieves 11.58% FPR95 on average, significantly outperforming PALM [40] by 3.38%. Further, for completeness, we extend the evaluation of our method to more strict settings, including (1) large-scale OOD detection, (2) hard OOD detection, and (3) unsupervised OOD detection.

## 2 Preliminary

**Notations.** We write vectors as bold-faced lowercase characters. Considering  $K$ -way classification as a case study, we use  $\mathcal{X}$  and  $\mathcal{Y} = \{1, \dots, K\}$  to indicate the input space and ID label space, respectively. The joint ID distribution, represented as  $P_{X_1 Y_1}$ , is a joint distribution defined over  $\mathcal{X} \times \mathcal{Y}$ . During testing time, there are some unknown OOD joint distributions  $P_{X_O Y_O}$  defined over  $\mathcal{X} \times \mathcal{Y}^c$ , where  $\mathcal{Y}^c$  is the complementary set of  $\mathcal{Y}$ . We also denote  $p(\mathbf{x})$  as the density of the ID marginal distribution  $P_{X_1}$ . According to [15, 16], OOD detection can be formally defined as follows:

**Problem 1 (OOD Detection)** *Given a labeled ID dataset  $\mathcal{D} = \{(\mathbf{x}_1, y_1), \dots, (\mathbf{x}_N, y_N)\}$  which is drawn from  $P_{X_1 Y_1}$  independent and identically distributed, the aim of OOD detection is to learn a*

*predictor  $g(\cdot)$  by using  $\mathcal{D}$  such that for any test data  $\mathbf{x}$ : 1) if  $\mathbf{x}$  is drawn from  $P_{X_1}$ , then  $g$  can classify  $\mathbf{x}$  into correct ID classes, and 2) if  $\mathbf{x}$  is drawn from  $P_{X_0}$ , then  $g$  can detect  $\mathbf{x}$  as OOD data.*

**OOD Scoring.** Existing methods [25, 62, 34, 67, 37] tend to adopt a post-hoc strategy to detect OOD samples, *i.e.*, given a well-trained discriminative model  $f_{\theta} : \mathcal{X} \rightarrow \mathbb{R}^d$  using  $\mathcal{D}$ , and a scoring function  $S$ , then  $\mathbf{x}$  is detected as ID data if and only if  $S(\mathbf{x}; f_{\theta}) \geq \lambda$ , for some given threshold  $\lambda$ :

$$g(\mathbf{x}) = \text{ID, if } S(\mathbf{x}; f_{\theta}) \geq \lambda; \text{ otherwise, } g(\mathbf{x}) = \text{OOD.} \quad (1)$$

A natural view for the motivation of the post-hoc strategy is to use a level set for ID density  $p(\mathbf{x})$  to discern ID and OOD data. To be specific, the main objective is to construct an efficient scoring function  $S$  that can effectively replicate the behaviour of the ID density function  $p(\mathbf{x})$  such that  $S(\mathbf{x}; f_{\theta}) \propto p(\mathbf{x})$ . From this perspective, let  $p_{\theta}(\mathbf{x})$  be the ID density function estimated by the discrimination model  $f_{\theta}$ , we can rewrite Eq. (1) as follow:

$$g(\mathbf{x}) = \text{ID, if } p_{\theta}(\mathbf{x}) \geq \lambda; \text{ otherwise, } g(\mathbf{x}) = \text{OOD.} \quad (2)$$

According to prior works, the design principle for  $p_{\theta}(\mathbf{x})$  can be either *logit-based* or *feature-based*.

**Logit-based OOD Methods** derive  $p_{\theta}(\mathbf{x})$  by formulating the discriminative model  $f_{\theta}$  as an energy-based model [18, 32], where a collection of energy values are turned into a probability density  $p_{\theta}(\mathbf{x})$  by implicitly resorting to the Gibbs-Boltzmann distribution, *i.e.*,

$$p_{\theta}(\mathbf{x}) = \frac{\exp[-E_{\theta}(\mathbf{z})/\tau]}{Z} \propto \exp[-E_{\theta}(\mathbf{z})/\tau], \quad E_{\theta}(\mathbf{z}) = -\tau \log \sum_{k=1}^K \exp(r_k/\tau), \quad (3)$$

where  $Z = \int \exp[-E_{\theta}(\mathbf{x})/\tau] d\mathbf{x}$  is an *input-independent* normalization constant and  $\mathbf{r} = [r_1, \dots, r_K] \in \mathbb{R}^K$  denotes the logit vector produced by a classification layer with the intermediate feature  $\mathbf{z} \in \mathbb{R}^d$  as the input.  $\tau > 0$  denotes a temperature hyperparameter and, when  $\tau \rightarrow 0$ , the negative energy score [37]  $-E_{\theta}(\cdot)$  will degenerate into the MaxLogit score [21, 78], namely,

$$\lim_{\tau \rightarrow 0} -E_{\theta}(\mathbf{z}) = \lim_{\tau \rightarrow 0} \tau \log \sum_{k=1}^K \exp(r_k/\tau) = \max_{k \in \mathcal{Y}} r_k.$$

**Feature-based OOD Methods** derive  $p_{\theta}(\mathbf{x})$  by explicitly assuming the intermediate feature space  $\mathcal{Z}$  learned by the discriminative model  $f_{\theta}$  to follow a pre-defined distribution. Therein, a representative example [48] is a Gaussian mixture distribution under a uniform ID-class prior (*i.e.*,  $p_{\theta}(k) = 1/K$ ). Formally, let  $\mathcal{N}(\boldsymbol{\mu}_k, \tau\Sigma)$  denotes the  $k$ -th ID class-conditional Gaussian distribution with a tied covariance matrix  $\tau\Sigma$ , then we have

$$p_{\theta}(\mathbf{x}) = \frac{1}{K} \sum_{k=1}^K \frac{\exp[-\frac{1}{2}(\mathbf{z} - \boldsymbol{\mu}_k)^{\top}(\tau\Sigma)^{-1}(\mathbf{z} - \boldsymbol{\mu}_k)]}{\sqrt{(2\pi)^d |\tau\Sigma|}} \propto \sum_{k=1}^K \exp\left[-\frac{1}{2}(\mathbf{z} - \boldsymbol{\mu}_k)^{\top}(\tau\Sigma)^{-1}(\mathbf{z} - \boldsymbol{\mu}_k)\right]. \quad (4)$$

Clearly, Eq. (4) is equivalent to the GEM score [34] when  $t = 1$ . Reducing  $\tau$  gradually reinforces the intra-class intensity and  $\tau \rightarrow 0$  degenerates Eq. (4) into the maximum Mahalanobis distance [34] (up to a constant), *i.e.*,

$$\begin{aligned} \lim_{\tau \rightarrow 0} \tau \log p_{\theta}(\mathbf{x}) &= \lim_{\tau \rightarrow 0} \tau \log \frac{1}{K} \sum_{k=1}^K \frac{\exp[-\frac{1}{2}(\mathbf{z} - \boldsymbol{\mu}_k)^{\top}(\tau\Sigma)^{-1}(\mathbf{z} - \boldsymbol{\mu}_k)]}{\sqrt{(2\pi)^d |\tau\Sigma|}} \\ &= \lim_{\tau \rightarrow 0} \tau \log \sum_{k=1}^K \exp\left[-\frac{1}{2}(\mathbf{z} - \boldsymbol{\mu}_k)^{\top} \Sigma^{-1}(\mathbf{z} - \boldsymbol{\mu}_k)/\tau\right] - \lim_{\tau \rightarrow 0} \tau \log K \sqrt{(2\pi)^d |\tau\Sigma|} \\ &= \max_{k \in \mathcal{Y}} -(\mathbf{z} - \boldsymbol{\mu}_k)^{\top} \Sigma^{-1}(\mathbf{z} - \boldsymbol{\mu}_k) - \underbrace{\lim_{\tau \rightarrow 0} \tau \log K \sqrt{(2\pi)^d |\tau\Sigma|}}_{\text{constant } \forall \mathbf{x} \in \mathcal{X}}. \end{aligned}$$

### 3 Methodology

#### 3.1 Motivation

We hereby present the motivation behind our work. In light of the aforementioned discussion, it can be found that existing OOD detection methods tend to make distributional assumptions on  $p_{\theta}(\mathbf{x})$

either explicitly or implicitly with the aim of density estimation due to the lacking true knowledge of  $p_\theta(\mathbf{x})$ . However, these distributional assumptions would not be held in many practical scenarios, leading to the mismatch between distributions of the learned and assumed features.

To illustrate this for the logit-based methods, we forge a mathematical connection between the classification objective and ID density function  $p_\theta(\mathbf{x})$ :

$$\log \frac{\exp(r_y/\tau)}{\sum_{k=1}^K \exp(r_k/\tau)} = r_y/\tau - \log \sum_{k=1}^K \exp(r_k/\tau) = r_y/\tau - \log Z p_\theta(\mathbf{x}), \quad (5)$$

where  $Z$  is the constant in Eq. 3. Maximizing the classification objective in Eq. 5 is to maximize the logit  $r_k$  while minimizing the estimated density function  $p_\theta(\mathbf{x})$ . This implies that training a discriminative model drives the ID marginal distribution away from the Gibbs-Boltzmann case. This cannot support the idea that the ID marginal distribution can necessarily align with the Gaussian distribution in the feature-based ones. Furthermore, it has been argued by [17] that the learned latent features fail the Henze-Zirkler multivariate normality test [24], which challenges the rationality of Gaussian-based distributional assumptions. We notice the possibility of non-parametric density estimation based on nearest-neighbor distance [17, 62] in the latent feature space. Albeit assumption-free, this practice hampers the scalability as it takes  $\mathcal{O}(Nd)$  memory to store the latent feature of ID training samples and  $\mathcal{O}(Nd)$  computation for neighborhood discovery. Most recently, [54] empirically shows that parametric density estimation can significantly benefit OOD detection more than the non-parametric counterpart if more reasonable distributional assumptions are introduced.

However, the bottleneck of OOD detection still lies in the consistency between the assumed distribution and the unknown ground truth. In view of this, this paper proposes to deterministically shape the learned feature distribution while training a discriminative model so that we can relax the widely adopted distributional assumptions.

### 3.2 Data distribution modeling

To preserve inter-class structures, this work, following common practice [13, 46], explicitly models the ID marginal distribution  $P_{X_1}$  as a weighted aggregation of ID class-conditioned distributions, *i.e.*,

$$\log p_\theta(\mathbf{x}) = \log \sum_{k=1}^K p_\theta(\mathbf{x}|k) \cdot p_\theta(k). \quad (6)$$

Consistent with probabilistic theory,  $p_\theta(\mathbf{x}|k)$  can be formulated into the following general form:

$$p_\theta(\mathbf{x}|k) = \frac{h(\mathbf{z}, k)}{\Phi(k)}, \quad \Phi(k) = \int_{\mathbf{z} \in \mathcal{Z}} h(\mathbf{z}, k) d\mathbf{z}, \quad (7)$$

where  $h(\mathbf{z}, k)$  represents a non-negative density function defined over the lower-dimensional latent feature space in which high-frequency and imperceptible details are abstracted away [55]. Note that our method is generic to the choice of  $h$  while this paper, inspired by [69], focuses on an exemplar based on the Euclidean distance, *i.e.*,  $h(\mathbf{z}, k) = \exp(-\|\mathbf{z} - \boldsymbol{\mu}_k\|_2^2/2\tau)$ , that is closely connected to universally optimal point configurations [4, 8]. Motivated by [13, 46], we further introduce  $\ell_2$  normalization over the latent feature space to keep  $\|\mathbf{z}\|_2 = 1$ . Without loss of generalization, let  $\hat{\boldsymbol{\mu}}_{kj} = \boldsymbol{\mu}_k / \|\boldsymbol{\mu}_k\|_2$  and  $\epsilon_k = \|\boldsymbol{\mu}_k\|_2 / \tau$ , we then have:

$$p_\theta(\mathbf{x}|k) = \frac{\exp\left[-(1 - 2 \cdot \hat{\boldsymbol{\mu}}_k^\top \mathbf{z} + \|\boldsymbol{\mu}_k\|_2^2)/2\tau\right]}{\int_{\mathbf{z}' \in \mathcal{Z}} \exp\left[-(1 - 2 \cdot \hat{\boldsymbol{\mu}}_k^\top \mathbf{z}' + \|\boldsymbol{\mu}_k\|_2^2)/2\tau\right] d\mathbf{z}'} = \frac{\exp(\epsilon_k \hat{\boldsymbol{\mu}}_k^\top \mathbf{z})}{\pi(\hat{\boldsymbol{\mu}}_k, \epsilon_k)}, \quad (8)$$

where  $\pi(\hat{\boldsymbol{\mu}}_k, \epsilon_k) = \int_{\mathbf{z} \in \mathcal{Z}} \exp(\epsilon_k \hat{\boldsymbol{\mu}}_k^\top \mathbf{z}) d\mathbf{z}$ .

### 3.3 Online approximation of normalization constant $\pi(\hat{\boldsymbol{\mu}}_k, \epsilon_k)$

Regarding the computation of  $p_\theta(\mathbf{x}|k)$  in Eq. (8), it is imperative to accurately calculate the normalization constant  $\pi(\hat{\boldsymbol{\mu}}_k, \epsilon_k)$  that seems to be intractable due to the integral over the latent feature space. When connecting Eq. (8) with the well-known von Mises-Fisher (vMF) distribution, one

can easily check that the normalization constant  $\pi(\hat{\mu}_k, \epsilon_k) = \epsilon_k^{d/2-1} / [(2\pi)^{d/2} I_{d/2-1}(\epsilon_k)]$  where  $I_p(\cdot)$  denotes the modified Bessel function of the first kind and order  $p$ . Unfortunately,  $I_p(\cdot)$  is a complicated function. To cast off this dilemma, the most straightforward idea, as in [13, 40, 46], involves fixing the  $\ell_2$  norm of  $\mu_k$  to be a constant value, *e.g.*,  $\|\mu_k\|_2 = 1, \forall k$ . This will result in  $\epsilon_k = 1/\tau, \forall k$  and therefore allows one to bypass the computation of the normalization constant  $\pi(\hat{\mu}_k, \epsilon_k)$  as  $\pi(\hat{\mu}_k, \epsilon_k)$  turns to be a class-independent constant term during optimization. Despite the simplicity, it is crucial to acknowledge that this practice implicitly implies the rigorous assumption that all classes ought to have a similar level of concentration, and would likely reduce the flexibility of the mixture model in Eq. (6). Indeed, this inflexibility in the latent feature space tends to deteriorate the learning of the feature mapping function. In this paper, we proceed from a different perspective and propose to use an online approximation thereof. To be precise, with a sufficiently large  $p$ ,  $I_p(\cdot)$  can be approximated using the following uniform expansion [39], *i.e.*,

$$I_p(pq) \sim \frac{\exp(p\nu)}{(2\pi p)^{1/2}(1+q^2)^{1/4}} \sum_{j=0}^{\infty} \frac{U_j(s)}{p^j}, \quad \nu = (1+q^2)^{1/2} + \log \frac{q}{1+(1+q^2)^{1/2}} \quad (9)$$

where  $\sim$  denotes a Poincaré asymptotic expansion with regard to  $p, s = (1+q^2)^{1/2}$  and polynomials  $U_j(s)$ . Empirically, we find that  $p = 127$ , which corresponds to the latent feature dimension  $d = 256$ , and evaluating only the first term in the sum, *i.e.*,  $U_0(s) = 1$ , are sufficiently enough to contribute to a satisfactory approximation while keeping simplicity. Please refer to Appendix A for more details.

### 3.4 DRL as Expectation-Maximization

In practice, it is hard to directly optimize the log-likelihood function [3, 64, 68] due to the existence of trivial solutions in which the latent feature space either scales up towards infinity or collapses to the origin point. To avoid collapse while making the classification process maximally beneficial to distribution modeling, let us start from the following evidence lower bound (ELBO) of  $\log p_{\theta}(\mathbf{x})$ :

$$\begin{aligned} \log p_{\theta}(\mathbf{x}) &= \sum_{k=1}^K q(k|\mathbf{x}) \log p_{\theta}(\mathbf{x}|k) - \text{KL}[q(k|\mathbf{x}) || p_{\theta}(k)] + \text{KL}[q(k|\mathbf{x}) || p_{\theta}(k|\mathbf{x})] \\ &\geq \sum_{k=1}^K q(k|\mathbf{x}) \log p_{\theta}(\mathbf{x}|k) - \text{KL}[q(k|\mathbf{x}) || p_{\theta}(k)] = \text{ELBO}(q, \mathbf{x}; \theta), \end{aligned} \quad (10)$$

where  $\text{KL}[\cdot]$  denotes the Kullback–Leibler (KL) divergence. We derive this ELBO in Appendix B. To make the inequality hold with equality so that the ELBO reaches its maximum value  $\log p(\mathbf{x})$ , we require  $\text{KL}[q(k|\mathbf{x}) || p_{\theta}(k|\mathbf{x})] = 0$ . By replacing  $q(k|\mathbf{x})$  with  $p_{\theta}(k|\mathbf{x})$ , we can formulate the maximization of the ELBO into the expectation–maximization (EM) framework.

**E-step.** With the fixed  $\theta_t$  at the iteration  $t$ , this step aims to estimate  $q_{t+1}(k|\mathbf{x})$  to make  $q_{t+1}(k|\mathbf{x}) = p_{\theta_t}(k|\mathbf{x})$  so that we can have  $\text{ELBO}(q_{t+1}, \mathbf{x}; \theta_t) = p_{\theta_t}(\mathbf{x})$ . Here, we estimate  $p_{\theta_t}(k|\mathbf{x})$  with the soft labels produced by the discrimination model at the iteration  $t$ , *i.e.*,  $f$  parameterized by  $\theta_t$ . To this end, by applying Bayes’ theorem to  $p_{\theta_t}(\mathbf{x})$ , we approximate  $q_{t+1}(k|\mathbf{x})$  as:

$$q_{t+1}(k|\mathbf{x}) = \frac{p_{\theta}(\mathbf{x}|k)p_{\theta}(k)}{\sum_{c=1}^K p_{\theta}(\mathbf{x}|c)p_{\theta}(c)} = \frac{\exp(\epsilon_y \hat{\mu}_y^\top \mathbf{z}) / \pi(\hat{\mu}_y, \epsilon_y)}{\sum_{k=1}^K \exp(\epsilon_k \hat{\mu}_k^\top \mathbf{z}) / \pi(\hat{\mu}_k, \epsilon_k)} \quad (11)$$

**M-step.** With the sub-optimal  $q_{t+1}(k|\mathbf{x}) = p_{\theta_t}(k|\mathbf{x})$  after E-step, we turn to maximize the ELBO. Given that  $p_{\theta}(k) = 1/K$  and  $q_{t+1}(k|\mathbf{x})$  is fixed during optimization, the KL term in  $\text{ELBO}(q, \mathbf{x}; \theta)$  can be reduced to a constant form  $-\text{KL}[q(k|\mathbf{x}) || p_{\theta}(k)] = \log K + H[q(k|\mathbf{x})]$ , which results in:

$$\theta_{t+1} = \arg \max_{\theta_t} \mathbb{E}_{(\mathbf{x}, y) \in \mathcal{D}} \left[ \sum_{k=1}^K q(k|\mathbf{x}) \log p_{\theta_t}(\mathbf{x}|k) \right]. \quad (12)$$

**Convergence.** At the E-step of the iteration  $t$ , we estimate  $q_{t+1}(k|\mathbf{x})$  to ensure  $\text{ELBO}(q_{t+1}, \mathbf{x}; \theta_t) = \log p_{\theta_t}(k|\mathbf{x})$ . At the M-step after the E-step, we have obtained  $\theta_{t+1}$  with a fixed  $q_{t+1}(k|\mathbf{x})$  such that  $\text{ELBO}(q_{t+1}, \mathbf{x}; \theta_{t+1}) \geq \text{ELBO}(q_{t+1}, \mathbf{x}; \theta_t)$ . Accordingly, we arrive at the following sequence:

$$\log p_{\theta_{t+1}}(\mathbf{x}) \geq \text{ELBO}(q_{t+1}, \mathbf{x}; \theta_{t+1}) \geq \text{ELBO}(q_{t+1}, \mathbf{x}; \theta_t) = \log p_{\theta_t}(\mathbf{x}). \quad (13)$$

Since  $\log p_{\theta_{t+1}}(k|\mathbf{x}) \geq \log p_{\theta_t}(k|\mathbf{x})$ , one can guarantee that  $\text{ELBO}(q, \mathbf{x}; \boldsymbol{\theta})$  is upper-bounded and can converge to a certain value with the proposed EM framework.

**Implementation.** When integrating the EM framework into the batch-based training routine, we note that the optimization of  $\text{ELBO}(q, \mathbf{x}; \boldsymbol{\theta})$  would suffer from the inconsistency issue. This is because the prediction  $p_{\theta}(k|\mathbf{x})$  only gets updated when  $\mathbf{x}$  is fed as the input. In other words, the prediction  $p_{\theta}(k|\mathbf{x})$  is only updated once per training epoch. However, the discriminative model  $f_{\theta}$  keeps updated throughout mini-batches in the training epoch. To address this problem, we rearrange the sequential sampling by constraining half of each mini-batch coinciding with the previous mini-batch. Meanwhile, the rest of the half will coincide with the next mini-batch. In this way, half of the current mini-batch can obtain on-the-fly soft labels generated by the discriminative model  $f_{\theta}$  in the previous mini-batch. For a clear notation, let  $\mathcal{B}_m = \{\mathcal{B}_m^{\text{pre}}, \mathcal{B}_m^{\text{next}}\}$  be the  $m$ -th mini-batch sampled from  $\mathcal{D}$  in the current training epoch, we then have  $\mathcal{B}_{m-1}^{\text{next}} = \mathcal{B}_m^{\text{pre}}$  and  $\mathcal{B}_1^{\text{pre}} = \emptyset$ .

### 3.5 Overall training objective

Our method enables end-to-end training, where the overall training objective  $\mathcal{R}(\cdot; \boldsymbol{\theta})$  is to maximize a linear combination of the classification objective  $\mathcal{R}_{\text{cls}}(\cdot, \cdot; \boldsymbol{\theta})$  and the ELBO in Eq. (10), *i.e.*,

$$\mathcal{R}(\mathcal{B}_m; \boldsymbol{\theta}) = \mathbb{E}_{(\mathbf{x}, y) \in \mathcal{B}_m} \mathcal{R}_{\text{cls}}(\mathbf{x}, y; \boldsymbol{\theta}) + \beta \mathbb{E}_{(\mathbf{x}, y) \in \mathcal{B}_m^{\text{pre}}} \text{ELBO}(q, \mathbf{x}; \boldsymbol{\theta}), \quad (14)$$

where the hyperparameter  $\beta > 0$  modulates the relative importance of two losses. Based on our proposed distribution modeling in Eq. (6), the classification for a sample  $\mathbf{x}$  can take place with a Bayes-based rule instead of a parametric softmax layer. As such, the the classification objective  $\mathcal{R}_{\text{cls}}(\mathbf{x}, y; \boldsymbol{\theta})$  naturally turns to be:

$$\mathcal{R}_{\text{cls}}(\mathbf{x}, y; \boldsymbol{\theta}) = \log \frac{\exp(\epsilon_y \hat{\boldsymbol{\mu}}_y^\top \mathbf{z}) / \pi(\hat{\boldsymbol{\mu}}_y, \epsilon_y)}{\sum_{k=1}^K \exp(\epsilon_k \hat{\boldsymbol{\mu}}_k^\top \mathbf{z}) / \pi(\hat{\boldsymbol{\mu}}_k, \epsilon_k)}. \quad (15)$$

**Remark.** While recent works [13, 40, 46] come with a Bayes-based classification objective as well by modeling the latent feature space as a mixture of vMF distributions. However, we emphasize that optimizing the classification objective alone does not necessarily drive the extracted training features towards the pre-defined distribution. For example, a latent feature  $\mathbf{z}$  can be far away from the corresponding class prototype  $\boldsymbol{\mu}_y$  while still being correctly classified as long as  $\mathbf{z}$  is relatively closer to  $\boldsymbol{\mu}_y$  than to other class prototypes. We, accordingly, mitigate this problem with the ELBO in Eq. (10) to explicitly measure the extent to which a training sample fits the assumed distribution. Besides, these methods are limited by assuming the concentration parameter  $\epsilon_k$  to be class-uniform for a simplified computation of Eq. (15) even though [13] considers optimizing  $\epsilon_k$  along with the discriminative model. By contrast, our method learns  $\epsilon_k$  directly through the prototype magnitudes while proposing an online approximation of the normalization constant  $\pi(\hat{\boldsymbol{\mu}}_k, \epsilon_k)$  to enable end-to-end training.

## 4 Related work

**Out-of-Distribution Detection** has attracted a surge of interest in recent years, which is motivated by the empirical observation [50] that neural networks tend to be over-confident in OOD data [52]. One line of work performs OOD detection by devising post-hoc scoring functions, including confidence-based methods [21, 38, 78], energy-based methods [37, 67], distance-based approaches [2, 34, 62, 36, 57, 76, 77], gradient-based approaches [25], and Bayesian approaches [29, 44]. Another line of work addresses OOD detection by fine-tuning a pre-trained discrimination model with training-time regularizations that help the model learn ID/OOD discrepancy following the guideline of outlier exposure [23] or negative prompts [53]. For instance, the discriminative model is regularized to produce lower confidence [33, 43], smaller feature magnitudes [37] or higher energy [10] for outlier points. More recently, some works have considered a practical scenario where the auxiliary outliers can be arbitrarily different from the real OOD data, therefore, distributionally augmenting the observed OOD data. Besides, the given OOD samples tend to include unlabelled ID counterparts [27]. In view of this, WOOD [27] formulates learning with noisy OOD samples as a constrained optimization problem while SAL [12] separates candidate outliers from the unlabeled data and then trains a binary classifier using the candidate outliers and the labeled ID data. Most regularization methods, unfortunately, assume the availability of auxiliary OOD data, not to mention resource-intensive re-training processes, while our method maintains the same training scheme as standard cross-entropy loss without requiring additional OOD data in training.

Table 1: OOD detection results on the CIFAR-10 benchmark with ResNet-18.  $\uparrow$  indicates larger values are better and vice versa. The best results in the last two columns are shown in bold.

Method	SVHN		Places365		LSUN		iSUN		Texture		Average	
	FPR95 $\downarrow$	AUROC $\uparrow$										
MSP	59.66	91.25	62.46	88.64	51.93	92.73	54.57	92.12	66.45	88.50	59.01	90.65
ODIN	20.93	95.55	63.04	86.57	31.92	94.82	33.17	94.65	56.40	86.21	41.09	91.56
Energy	54.41	91.22	37.22	92.70	10.19	98.05	27.52	95.59	55.23	89.37	36.91	93.39
ReAct	48.16	92.32	37.25	93.13	18.09	96.91	20.35	95.59	96.51	47.41	34.25	94.09
ASH	28.94	94.84	27.29	91.31	9.06	98.34	21.61	95.95	35.02	93.63	27.29	94.81
Maha	9.24	97.80	83.50	69.56	67.73	73.61	6.02	98.63	23.21	92.91	37.94	86.50
KNN	27.97	95.48	47.84	89.93	18.50	96.84	24.68	95.52	26.74	94.96	29.15	94.55
CONJ	18.71	96.48	53.44	89.18	22.20	95.95	22.64	95.87	25.80	95.11	28.56	94.52
Vim	24.95	95.36	63.04	86.57	7.26	98.53	33.17	94.65	56.40	86.21	36.96	92.26
VOS	15.69	96.37	37.95	91.78	27.64	93.82	30.42	94.87	32.68	93.68	28.88	94.10
CSI	37.38	94.69	38.31	93.04	10.63	97.93	10.36	98.01	28.85	94.87	25.11	95.71
SSD+	2.47	99.51	22.05	95.57	10.56	97.83	28.44	95.67	9.27	98.35	14.56	97.38
KNN+	2.70	99.61	23.05	94.88	7.89	98.01	24.56	96.21	10.11	97.43	13.66	97.22
CIDER	2.89	99.72	23.05	94.09	5.45	99.01	20.21	96.64	12.33	96.85	12.95	97.26
PALM	0.34	99.91	28.81	94.80	1.11	99.65	34.07	95.17	10.48	98.29	14.96	97.57
DRL	7.91	98.82	19.17	95.65	12.87	99.09	11.92	98.12	4.92	97.48	<b>11.58</b>	<b>97.83</b>

**Representation Learning for OOD Detection** is an emerging topic to enhance the performance of post-hoc distance-based OOD detectors. In particular, CSI [63] investigates the type of data augmentations that are particularly beneficial for OOD detection while other works [58, 70] verify the effectiveness of applying the off-the-shelf multi-view contrastive losses such as SimCLR [5] and SupCon [28] for OOD detection. CIDER [46] proposes a prototypical contrastive learning framework for OOD detection by promoting stronger ID-OOD separability than SupCon loss, where a regularization strategy is to ensure that all samples are compactly located around their corresponding class prototype. PALM [40] extends CIDER by introducing a mixture of prototypes to represent each class and performing prototype-level contrastive learning to enhance intra-class compactness and inter-class discrimination. This paper significantly differs from PALM and CIDER in the following two aspects: 1) we avoid normalizing the class prototypes to protect the generalization ability of the data distributional modeling, and 2) we learn a feature space that best describes the pre-defined distribution rather than features that preserve inter- and intra- class structures.

## 5 Experiments

**Software and Hardware.** We perform all experiments on an NVIDIA A100 GPU using Pytorch.

**Baseline Methods.** We compare DRL with representative methods, including MSP [22], ODIN [35], Energy [37], ReAct [37], ASH [11], Mahalanobis (Maha) [34], KNN [62], CONJ [54], Vim [66], VOS [14], CSI [63], SSD+ [58], KNN [37], CIDER [46], and PALM [40]. It is worth noting that we have adopted the recommended configurations proposed by prior works, while concurrently standardizing the backbone architecture to ensure equitable comparisons.

**Evaluation Metrics.** The detection performance is evaluated via three widely used metrics: 1) the false positive rate of OOD data is measured when the true positive rate of ID data reaches 95% (FPR95); 2) the area under the receiver operating characteristic curve (AUROC) is computed to quantify the probability of the ID case receiving a higher score than the OOD case. The reported results of DRL are averaged over 5 independent runs.

### 5.1 Main results on CIFAR benchmarks

Following the setup in [46, 40], we consider CIFAR-10 [30] and CIFAR-100 [30] as ID datasets and train ResNet-18 [20] and ResNet-34 [20] on them respectively. We train the model using stochastic gradient descent with momentum 0.9, and weight decay  $10^{-4}$  for 500 epochs. The initial learning rate is 0.5 with cosine scheduling and the batch size is 512. There are six datasets for OOD detection with regard to CIFAR benchmarks: SVHN [49], LSUN [75], iSUN [73], Places [82], and Textures [6]. At inference time, all images are of size  $32 \times 32$ . Tables 1 and 2 presented the performance of different methods, where our method significantly outperforms existing methods. Specifically, compared with the mostly advanced PALM, DRL reveals 3.38% average improvements w.r.t. FPR95 on CIFAR-10.

Table 2: OOD detection results on the CIFAR-100 benchmark with ResNet-34.  $\uparrow$  indicates larger values are better and vice versa. The best results in the last two columns are shown in bold.

Method	SVHN		Places365		LSUN		iSUN		Texture		Average	
	FPR95 $\downarrow$	AUROC $\uparrow$										
MSP	78.89	79.80	84.38	74.21	83.47	75.28	84.61	74.51	86.51	72.53	83.12	75.27
ODIN	70.16	84.88	82.16	75.19	76.36	80.10	79.54	79.16	85.28	75.23	78.70	79.11
Energy	66.91	85.25	81.41	76.37	59.77	86.69	66.52	84.49	79.01	79.96	70.72	82.55
ReAct	50.93	88.75	83.55	73.10	64.02	80.31	81.80	79.99	64.40	81.95	68.94	80.82
ASH	52.96	90.19	72.62	76.38	75.18	76.52	55.55	87.86	56.17	86.75	62.50	83.53
Maha	87.09	80.62	84.63	73.89	84.15	79.43	83.18	78.83	61.72	84.87	80.15	79.53
KNN	46.25	90.39	82.08	75.44	60.85	85.61	71.56	86.28	62.39	83.95	64.63	84.33
CONJ	46.19	90.44	80.81	75.83	60.45	85.90	64.62	87.77	62.13	83.77	62.84	84.74
Vim	73.42	84.62	85.34	69.34	86.96	69.74	85.35	73.16	74.56	76.23	81.13	74.62
VOS	43.24	82.80	76.85	78.63	73.61	84.69	69.65	86.32	57.57	87.31	64.18	83.95
CSI	44.53	92.65	79.08	76.27	75.58	83.78	76.62	84.98	61.61	86.47	67.48	84.83
SSD+	31.19	94.19	77.74	79.90	79.39	85.18	80.85	84.08	66.63	86.18	67.16	85.91
KNN+	39.23	92.78	80.74	77.58	48.99	89.30	74.99	82.69	57.15	88.35	60.22	86.14
CIDER	23.09	95.16	79.63	73.43	16.16	96.33	71.68	82.98	43.87	90.42	46.89	87.67
PALM	3.29	99.23	64.66	84.72	9.86	98.01	28.71	94.64	33.56	92.49	<b>28.02</b>	<b>93.82</b>
DRL	20.15	94.07	76.64	77.55	16.97	94.63	32.57	92.33	31.97	92.09	35.66	90.13

## 5.2 Extensions

**Hard OOD Detection.** We consider hard OOD scenarios, in which the OOD data are semantically similar to those of the ID cases. With the CIFAR-100 as the ID dataset for training ResNet-34, we evaluate our method on 4 hard OOD datasets, namely, LSUN-Fix [75], ImageNet-Fix [9], ImageNet-Resize [9], and CIFAR-10. We select a set of strong baselines that are competent in hard OOD detection with results in Table 3. It can be found that our method can beat the state-of-the-art across the considered datasets, even for the challenging CIFAR-100 versus CIFAR-10 setting.

Table 3: Evaluation on hard OOD detection tasks.  $\uparrow$  indicates larger values are better and vice versa. The best result in each column is shown in bold.

Method	LSUN-Fix		ImageNet-Fix		ImageNet-Resize		CIFAR-10		Average	
	FPR95 $\downarrow$	AUROC $\uparrow$								
SSD+	83.36	76.63	76.73	79.78	83.67	81.09	85.16	73.70	82.23	77.80
KNN+	84.96	75.37	75.52	79.95	68.49	84.91	84.12	75.91	78.27	79.04
CIDER	90.94	70.31	78.83	77.53	56.89	87.62	84.87	73.30	77.88	77.19
PALM	77.15	77.24	66.19	82.51	27.02	95.03	87.25	72.28	64.40	81.76
DRL	68.64	78.75	59.92	87.19	40.85	91.11	74.12	80.93	<b>60.88</b>	<b>84.50</b>

**Unsupervised OOD Detection.** To verify the reliance of our method on the availability of ground-truth labels, we consider unsupervised OOD scenarios. Following [40], we take unlabelled CIFAR-100 as the ID dataset to train ResNet-34 from scratch. Due to the lack of ground-truth labels, we resort to maintaining a momentum teacher, inspired by DINO [83], to produce soft pseudo-labels as a surrogate. For a fair comparison, we use the same data augmentation techniques with [40].

Table 4: Evaluation on unsupervised OOD detection tasks.  $\uparrow$  indicates larger values are better and vice versa. The best results in the last two columns are shown in bold.

Method	SVHN		Places365		LSUN		iSUN		Texture		Average	
	FPR95 $\downarrow$	AUROC $\uparrow$										
KNN	61.21	84.92	81.46	72.97	69.65	77.77	93.35	70.39	78.49	76.75	74.83	76.56
SSD	60.13	86.40	79.05	73.68	61.94	84.47	84.37	75.58	71.91	83.35	71.48	80.70
CSI	14.47	97.14	86.23	66.93	34.12	94.21	87.79	80.15	80.15	92.13	53.55	86.11
PALM	13.86	97.53	85.63	69.46	21.28	95.95	53.43	89.06	42.62	88.33	43.37	88.07
DRL	25.11	94.50	79.06	74.10	36.88	90.88	30.66	93.96	31.59	91.58	<b>40.66</b>	<b>89.00</b>

**Large-scale OOD Detection.** We conduct experiments on the ImageNet benchmark, demonstrating the scalability of our method. Specifically, we inherit the setup from [40, 46], where the ID dataset is ImageNet-100 [9], and OOD datasets include iNaturalist [65], SUN [71], Places365 [83], and

Textures [7]. Following the setting in the previous works [40, 46], we fine-tune the last residual block of ResNet-50 [20] pre-trained on ImageNet-1K [9] for 20 epochs with the learning rate  $10^{-3}$  while freezing the rest of the parameters. At inference time, all images are resized to  $224 \times 224$ . In Figure 1, we reported the performances of four OOD test datasets respectively. It can be seen that our method reaches state-of-the-art results on average across four OOD datasets.

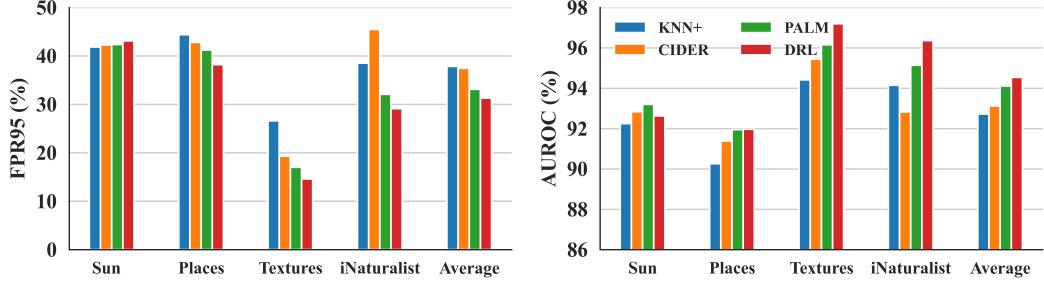


Figure 1: OOD detection results on the ImageNet-100 benchmark with ResNet-50.

### 5.3 Ablation study

In this section, we conduct an ablation study to validate our motivation and design. In Figure 2a, we compare the effect between  $\ell_2$ -normalized and unnormalized class prototypes. While  $\ell_2$ -normalized class prototypes enable DRL to avoid estimating the normalization constant, we can observe that unnormalized class prototypes come with a significant performance increase in detecting OOD samples. We suspect that unnormalized class prototypes come with a more reliable assumption that all classes are allowed to have a different level of concentration from each other. Besides, recalling that we have rearranged sequential sampling to optimize the ELBO of DRL in a consistent manner, Figure 2b empirically examines the effectiveness of this practice by demonstrating that the OOD detection performance considerably drops without the rearranging of sequential sampling.

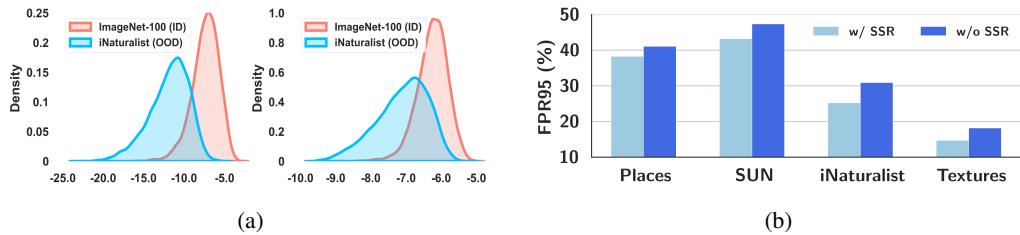


Figure 2: Ablation Study on the proposed DRL: (a) distribution of feature embeddings learned with unnormalized (left) and  $\ell_2$ -normalized (right) class prototypes; (b) the benefits of the proposed sequential sampling rearrange to OOD detection performance.

## 6 Conclusion

In this paper, we propose a novel learning framework DRL that mitigates the gap between pre-trained discriminative models and density-based post-hoc OOD detection. Methodologically, DRL focuses on answering an important yet under-explored question of whether it is possible to deterministically shape the ID feature distribution while pre-training a discriminative model. Theoretically, DRL is formulated as an Expectation-Maximization algorithm, where we design a bounded ELBO and rearrange the sequential sampling for consistent optimization. Empirically, DRL achieves consistently strong performance of OOD detection compared to competitive baselines on multiple benchmarks, which implies the superiority of our proposed DRL. We hope our work can inspire future research on shaping ID feature space for density-based post-hoc OOD detection.

## 7 Acknowledgement

This work was supported in part by the NSFC / Research Grants Council (RGC) Joint Research Scheme under the grant N\_HKBU214/21, the General Research Fund of RGC under the grants 12201321, 12202622, 12201323, and the RGC Senior Research Fellow Scheme under the grant: SRFS2324-2S02.

## References

- [1] Yong Hyun Ahn, Gyeong-Moon Park, and Seong Tae Kim. Line: Out-of-distribution detection by leveraging important neurons. In *CVPR*, 2023.
- [2] Mouin Ben Ammar, Nacim Belkhir, Sebastian Popescu, Antoine Manzanera, and Gianni Franchi. Neco: Neural collapse based out-of-distribution detection. In *ICLR*, 2024.
- [3] Christopher M Bishop. Pattern recognition and machine learning. *Springer Google Schola*, 2:645–678, 2006.
- [4] Sergiy V Borodachov, Douglas P Hardin, and Edward B Saff. *Discrete energy on rectifiable sets*, volume 4. Springer, 2019.
- [5] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *ICML*, 2020.
- [6] Mircea Cimpoi, Subhransu Maji, Iasonas Kokkinos, Sammy Mohamed, and Andrea Vedaldi. Describing textures in the wild. In *CVPR*, 2014.
- [7] Mircea Cimpoi, Subhransu Maji, Iasonas Kokkinos, Sammy Mohamed, and Andrea Vedaldi. Describing textures in the wild. In *CVPR*, 2014.
- [8] Henry Cohn and Abhinav Kumar. Universally optimal distribution of points on spheres. *Journal of the American Mathematical Society*, 20(1):99–148, 2007.
- [9] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *CVPR*, 2009.
- [10] Akshay Raj Dhamija, Manuel Günther, and Terrance Boult. Reducing network agnostophobia. In *NeurIPS*, 2018.
- [11] Andrija Djurisic, Nebojsa Bozanic, Arjun Ashok, and Rosanne Liu. Extremely simple activation shaping for out-of-distribution detection. In *ICLR*, 2023.
- [12] Xuefeng Du, Zhen Fang, Ilias Diakonikolas, and Yixuan Li. How does wild data provably help ood detection? In *ICLR*, 2024.
- [13] Xuefeng Du, Gabriel Gozum, Yifei Ming, and Yixuan Li. Siren: Shaping representations for detecting out-of-distribution objects. In *NeurIPS*, 2022.
- [14] Xuefeng Du, Zhaoning Wang, Mu Cai, and Sharon Li. Towards unknown-aware learning with virtual outlier synthesis. In *ICLR*, 2022.
- [15] Zhen Fang, Yixuan Li, Feng Liu, Bo Han, and Jie Lu. On the learnability of out-of-distribution detection. *Journal of Machine Learning Research*, 25, 2024.
- [16] Zhen Fang, Yixuan Li, Jie Lu, Jiahua Dong, Bo Han, and Feng Liu. Is out-of-distribution detection learnable? In *NeurIPS*, 2022.
- [17] Soumya Suvra Ghosal, Yiyou Sun, and Yixuan Li. How to overcome curse-of-dimensionality for out-of-distribution detection? In *AAAI*, 2024.
- [18] Will Grathwohl, Kuan-Chieh Wang, Jörn-Henrik Jacobsen, David Duvenaud, Mohammad Norouzi, and Kevin Swersky. Your classifier is secretly an energy based model and you should treat it like one. In *ICLR*, 2020.
- [19] Michael U Gutmann and Aapo Hyvärinen. Noise-contrastive estimation of unnormalized statistical models, with applications to natural image statistics. *Journal of machine learning research*, 13(2), 2012.
- [20] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, 2016.
- [21] Dan Hendrycks, Steven Basart, Mantas Mazeika, Andy Zou, Joe Kwon, Mohammadreza Mostajabi, Jacob Steinhardt, and Dawn Song. Scaling out-of-distribution detection for real-world settings. In *ICML*, 2022.
- [22] Dan Hendrycks and Kevin Gimpel. A baseline for detecting misclassified and out-of-distribution examples in neural networks. In *ICLR*, 2022.
- [23] Dan Hendrycks, Mantas Mazeika, and Thomas Dietterich. Deep anomaly detection with outlier exposure. In *ICLR*, 2019.
- [24] Norbert Henze and Bernd Zirkler. A class of invariant consistent tests for multivariate normality. *Communications in statistics-Theory and Methods*, 19(10):3595–3617, 1990.
- [25] Rui Huang, Andrew Geng, and Yixuan Li. On the importance of gradients for detecting distributional shifts in the wild. In *NeurIPS*, pages 677–689, 2021.
- [26] 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.
- [27] Julian Katz-Samuels, Julia B Nakhleh, Robert Nowak, and Yixuan Li. Training ood detectors in their natural habitats. In *ICML*, 2022.

[28] Prannay Khosla, Piotr Teterwak, Chen Wang, Aaron Sarna, Yonglong Tian, Phillip Isola, Aaron Maschinot, Ce Liu, and Dilip Krishnan. Supervised contrastive learning. In *NeurIPS*, 2020.

[29] Agustinus Kristiadi, Matthias Hein, and Philipp Hennig. Being bayesian, even just a bit, fixes overconfidence in relu networks. In *ICML*, 2020.

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

[31] Hao Lang, Yinhe Zheng, Yixuan Li, Jian Sun, Fei Huang, and Yongbin Li. A survey on out-of-distribution detection in nlp. *Transactions on Machine Learning Research*, 2023.

[32] Yann LeCun, Sumit Chopra, Raia Hadsell, M Ranzato, and Fujie Huang. A tutorial on energy-based learning. *Predicting Structured Data*, 1(0), 2006.

[33] Kimin Lee, Honglak Lee, Kibok Lee, and Jinwoo Shin. Training confidence-calibrated classifiers for detecting out-of-distribution samples. In *ICLR*, 2018.

[34] Kimin Lee, Kibok Lee, Honglak Lee, and Jinwoo Shin. A simple unified framework for detecting out-of-distribution samples and adversarial attacks. In *NeurIPS*, 2018.

[35] Shiyu Liang, Yixuan Li, and Rayadurgam Srikanth. Enhancing the reliability of out-of-distribution image detection in neural networks. In *ICLR*, 2018.

[36] Litian Liu and Yao Qin. Detecting out-of-distribution through the lens of neural collapse. In *ICLR*, 2024.

[37] Weitang Liu, Xiaoyun Wang, John Owens, and Yixuan Li. Energy-based out-of-distribution detection. In *NeurIPS*, 2020.

[38] Xixi Liu, Yaroslava Lochman, and Christopher Zach. Gen: Pushing the limits of softmax-based out-of-distribution detection. In *CVPR*, 2023.

[39] Daniel W Lozier. Nist digital library of mathematical functions. *Annals of Mathematics and Artificial Intelligence*, 38:105–119, 2003.

[40] Haodong Lu, Dong Gong, Shuo Wang, Jason Xue, Lina Yao, and Kristen Moore. Learning with mixture of prototypes for out-of-distribution detection. In *ICLR*, 2024.

[41] Jie Lu, Junyu Xuan, Guangquan Zhang, and Xiangfeng Luo. Structural property-aware multilayer network embedding for latent factor analysis. *Pattern Recognition*, 76:228–241, 2018.

[42] Jie Lu, Hua Zuo, and Guangquan Zhang. Fuzzy multiple-source transfer learning. *IEEE Transactions on Fuzzy Systems*, 28(12):3418–3431, 2019.

[43] Andrey Malinin and Mark Gales. Predictive uncertainty estimation via prior networks. In *NeurIPS*, 2018.

[44] Andrey Malinin and Mark Gales. Reverse kl-divergence training of prior networks: Improved uncertainty and adversarial robustness. In *NeurIPS*, 2019.

[45] Marc Masana, Xialei Liu, Bartłomiej Twardowski, Mikel Menta, Andrew D Bagdanov, and Joost Van De Weijer. Class-incremental learning: survey and performance evaluation on image classification. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(5):5513–5533, 2022.

[46] Yifei Ming, Yiyou Sun, Ousmane Dia, and Yixuan Li. How to exploit hyperspherical embeddings for out-of-distribution detection? In *ICLR*, 2023.

[47] Todd K Moon. The expectation-maximization algorithm. *IEEE Signal Processing Magazine*, 13(6):47–60, 1996.

[48] Peyman Morteza and Yixuan Li. Provable guarantees for understanding out-of-distribution detection. In *AAAI*, 2022.

[49] Yuval Netzer, Tao Wang, Adam Coates, Alessandro Bissacco, Bo Wu, and Andrew Y Ng. Reading digits in natural images with unsupervised feature learning. 2011.

[50] Anh Nguyen, Jason Yosinski, and Jeff Clune. Deep neural networks are easily fooled: High confidence predictions for unrecognizable images. In *CVPR*, 2015.

[51] Andre T Nguyen, Fred Lu, Gary Lopez Munoz, Edward Raff, Charles Nicholas, and James Holt. Out of distribution data detection using dropout bayesian neural networks. In *AAAI*, volume 36, pages 7877–7885, 2022.

[52] Jun Nie, Yadan Luo, Shanshan Ye, Yonggang Zhang, Xinmei Tian, and Zhen Fang. Out-of-distribution detection with virtual outlier smoothing. *International Journal of Computer Vision*, pages 1–18, 2024.

[53] Jun Nie, Yonggang Zhang, Zhen Fang, Tongliang Liu, Bo Han, and Xinmei Tian. Out-of-distribution detection with negative prompts. In *ICLR*, 2024.

[54] Bo Peng, Yadan Luo, Yonggang Zhang, Yixuan Li, and Zhen Fang. Conjnorm: Tractable density estimation for out-of-distribution detection. In *ICLR*, 2024.

[55] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *CVPR*, 2022.

[56] Mohammadreza Salehi, Hossein Mirzaei, Dan Hendrycks, Yixuan Li, Mohammad Hossein Rohban, and Mohammad Sabokrou. A unified survey on anomaly, novelty, open-set, and out-of-distribution detection: Solutions and future challenges. *Transactions on Machine Learning Research*, 2022.

[57] Chandramouli Shama Sastry and Sageev Oore. Detecting out-of-distribution examples with gram matrices. In *ICML*, 2020.

[58] Vikash Sehwag, Mung Chiang, and Prateek Mittal. Ssd: A unified framework for self-supervised outlier detection. In *ICLR*, 2021.

[59] Yue Song, Nicu Sebe, and Wei Wang. Rankfeat: Rank-1 feature removal for out-of-distribution detection. In *NeurIPS*, 2022.

[60] Yiyou Sun, Chuan Guo, and Yixuan Li. React: Out-of-distribution detection with rectified activations. In *NeurIPS*, 2021.

[61] Yiyou Sun and Yixuan Li. Dice: Leveraging sparsification for out-of-distribution detection. In *ECCV*, 2022.

[62] Yiyou Sun, Yifei Ming, Xiaojin Zhu, and Yixuan Li. Out-of-distribution detection with deep nearest neighbors. In *ICML*, 2022.

[63] Jihoon Tack, Sangwoo Mo, Jongheon Jeong, and Jinwoo Shin. Csi: Novelty detection via contrastive learning on distributionally shifted instances. In *NeurIPS*, 2020.

[64] Tsung Wei Tsai, Chongxuan Li, and Jun Zhu. Mice: Mixture of contrastive experts for unsupervised image clustering. In *ICLR*, 2020.

[65] Grant Van Horn, Oisin Mac Aodha, Yang Song, Yin Cui, Chen Sun, Alex Shepard, Hartwig Adam, Pietro Perona, and Serge Belongie. The inaturalist species classification and detection dataset. In *CVPR*, 2018.

[66] Haoqi Wang, Zhizhong Li, Litong Feng, and Wayne Zhang. Vim: Out-of-distribution with virtual-logit matching. In *CVPR*, 2022.

[67] Haoran Wang, Weitang Liu, Alex Bocchieri, and Yixuan Li. Can multi-label classification networks know what they don't know? In *NeurIPS*, 2021.

[68] Qizhou Wang, Bo Han, Tongliang Liu, Gang Niu, Jian Yang, and Chen Gong. Tackling instance-dependent label noise via a universal probabilistic model. In *AAAI*, 2021.

[69] Tongzhou Wang and Phillip Isola. Understanding contrastive representation learning through alignment and uniformity on the hypersphere. In *ICML*, 2020.

[70] Jim Winkens, Rudy Bunel, Abhijit Guha Roy, Robert Stanforth, Vivek Natarajan, Joseph R Ledsam, Patricia MacWilliams, Pushmeet Kohli, Alan Karthikesalingam, Simon Kohl, et al. Contrastive training for improved out-of-distribution detection. In *arXiv*, 2020.

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

[72] Mingyu Xu, Zheng Lian, Bin Liu, and Jianhua Tao. Vra: Variational rectified activation for out-of-distribution detection. In *NeurIPS*, 2023.

[73] Pingmei Xu, Krista A Ehinger, Yinda Zhang, Adam Finkelstein, Sanjeev R Kulkarni, and Jianxiong Xiao. Turkergaze: Crowdsourcing saliency with webcam based eye tracking. In *arXiv*, 2015.

[74] Jingkang Yang, Kaiyang Zhou, Yixuan Li, and Ziwei Liu. Generalized out-of-distribution detection: A survey. *International Journal of Computer Vision*, pages 1–28, 2024.

[75] 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. In *arXiv*, 2015.

[76] Yeonguk Yu, Sungho Shin, Seongju Lee, Changhyun Jun, and Kyoobin Lee. Block selection method for using feature norm in out-of-distribution detection. In *CVPR*, 2023.

[77] Jinsong Zhang, Qiang Fu, Xu Chen, Lun Du, Zelin Li, Gang Wang, Shi Han, Dongmei Zhang, et al. Out-of-distribution detection based on in-distribution data patterns memorization with modern hopfield energy. In *ICLR*, 2022.

[78] Zihan Zhang and Xiang Xiang. Decoupling maxlogit for out-of-distribution detection. In *CVPR*, 2023.

[79] Qinyu Zhao, Ming Xu, Kartik Gupta, Akshay Asthana, Liang Zheng, and Stephen Gould. Towards optimal feature-shaping methods for out-of-distribution detection. In *ICLR*, 2024.

[80] Qinyu Zhao, Ming Xu, Kartik Gupta, Akshay Asthana, Liang Zheng, and Stephen Gould. Towards optimal feature-shaping methods for out-of-distribution detection. In *ICLR*, 2024.

[81] Zhong-Qiu Zhao, Peng Zheng, Shou-tao Xu, and Xindong Wu. Object detection with deep learning: A review. *IEEE Transactions on Neural Networks and Learning Systems*, 30(11):3212–3232, 2019.

[82] Bolei Zhou, Agata Lapedriza, Aditya Khosla, Aude Oliva, and Antonio Torralba. Places: A 10 million image database for scene recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 40(6):1452–1464, 2017.

[83] Bolei Zhou, Agata Lapedriza, Aditya Khosla, Aude Oliva, and Antonio Torralba. Places: A 10 million image database for scene recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 40(6):1452–1464, 2017.

[84] Yao Zhu, YueFeng Chen, Chuanlong Xie, Xiaodan Li, Rong Zhang, Hui Xue, Xiang Tian, Yaowu Chen, et al. Boosting out-of-distribution detection with typical features. In *NeurIPS*, 2022.

[85] David Zimmerer, Peter M Full, Fabian Isensee, Paul Jäger, Tim Adler, Jens Petersen, Gregor Köhler, Tobias Ross, Anniika Reinke, Antanas Kascenas, et al. Mood 2020: A public benchmark for out-of-distribution detection and localization on medical images. *IEEE Transactions on Medical Imaging*, 41(10):2728–2738, 2022.

## A Approximating vMF normalization constant

To approach the vMF probability density function for the random  $d$ -dimensional unit vector and concentration parameter  $\epsilon$ , one needs to get the normalization constant by calculating  $\epsilon^{d/2-1} / [(2\pi)^{d/2} I_{d/2-1}(\epsilon)]$  where  $I_p(\cdot)$  denotes the modified Bessel function of the first kind and order  $p$ . However,  $I_p(\cdot)$  is a complicated function. To cast off this dilemma, as explained by Eq. (9) in the main paper, we propose to approximate  $I_p(\cdot)$  by resorting to an asymptotic expansion. We compare the values of  $\log \hat{I}_p(\cdot)$  obtained by our proposed approximation with the values of  $\log I_p(\cdot)$  obtained by the scipy implementation<sup>2</sup>, where we define an approximation error  $\delta(\epsilon)$  as follows:

$$\delta(\epsilon) = \log \hat{I}_p(\epsilon) - \log I_p(\epsilon) \quad (16)$$

It can be seen from Figure 3 that the approximation error is small compared to the actual function values when  $p = 127$ . This implies that the latent feature dimension  $d = 256$  is sufficiently large to get a good approximation of the normalization constant.

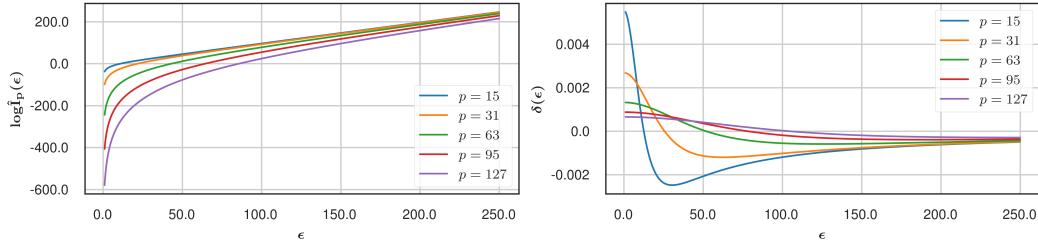


Figure 3: Left: The actual value of our proposed approximation of  $\log \hat{I}_p(\epsilon)$ ; Right: The approximation error  $\delta(\epsilon)$  in computing  $I_p(\cdot)$  between our proposed and the scipy implementation.

## B Training stability on CIFAR-100

To verify that our method consistently provides strong performance, we train with 5 independent seeds for CIFAR-100 and report the average and standard deviation of FPR and AUROC in Table 5.

Table 5: Ablation on stability. OOD detection performance of our method on CIFAR- 100. Results are averaged over 5 independent runs.

Method	SVHN		Places365		LSUN		iSUN		Texture	
	FPR95↓	AUROC↑	FPR95↓	AUROC↑	FPR95↓	AUROC↑	FPR95↓	AUROC↑	FPR95↓	AUROC↑
Mean	10.15	98.07	66.64	81.55	6.97	98.63	22.57	96.33	21.97	96.09
Std.	1.52	0.11	3.06	2.41	2.28	1.46	3.25	0.87	2.24	1.94

## C Limitations

This paper only explores one type of realization of the non-negative density function. It will be exciting to explore more possibilities for the realization and parameterization.

## D Broader impacts

Our project aims to improve the reliability and safety of modern machine learning models. Our study can lead to direct benefits and societal impacts, particularly for safety-critical applications such as autonomous driving. Our study does not involve any human subjects or violation of legal compliance. We do not anticipate any potentially harmful consequences to our work. Through our study, we hope to raise stronger research and societal awareness towards the problem of out-of-distribution detection in real-world settings.

<sup>2</sup><https://docs.scipy.org/doc/scipy/reference/generated/scipy.special.iv.html>

## NeurIPS Paper Checklist

### 1. Claims

Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

Answer: **[Yes]**

Justification: We clearly illustrate the focused problems and state our contribution methodologically, theoretically and empirically in the abstract and introduction.

Guidelines:

- The answer NA means that the abstract and introduction do not include the claims made in the paper.
- The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
- The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
- It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

### 2. Limitations

Question: Does the paper discuss the limitations of the work performed by the authors?

Answer: **[Yes]**

Justification: We discuss the limitation of our method in the appendix.

Guidelines:

- The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
- The authors are encouraged to create a separate "Limitations" section in their paper.
- The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
- The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
- The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
- The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
- If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
- While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

### 3. Theory Assumptions and Proofs

Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

Answer: **[NA]**

Justification: the paper does not include theoretical results

Guidelines:

- The answer NA means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

#### 4. Experimental Result Reproducibility

Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

Answer: **[Yes]**

Justification: We provide implementation details of our method in the Experiment section and the appendix.

Guidelines:

- The answer NA means that the paper does not include experiments.
- If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
- If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
- Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general, releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
- While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
  - (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
  - (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
  - (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
  - (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

#### 5. Open access to data and code

Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

Answer: [No]

Justification: We will release our code upon acceptance.

Guidelines:

- The answer NA means that paper does not include experiments requiring code.
- Please see the NeurIPS code and data submission guidelines (<https://nips.cc/public/guides/CodeSubmissionPolicy>) for more details.
- While we encourage the release of code and data, we understand that this might not be possible, so “No” is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (<https://nips.cc/public/guides/CodeSubmissionPolicy>) for more details.
- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

## 6. Experimental Setting/Details

Question: Does the paper specify all the training and test details (e.g., data splits, hyper-parameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

Answer: [Yes]

Justification: We provide implementation details of our method in the Experiment section.

Guidelines:

- The answer NA means that the paper does not include experiments.
- The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
- The full details can be provided either with the code, in appendix, or as supplemental material.

## 7. Experiment Statistical Significance

Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

Answer: [Yes]

Justification: We have provided the standard deviation of our method in the appendix.

Guidelines:

- The answer NA means that the paper does not include experiments.
- The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.

- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
- If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

## 8. Experiments Compute Resources

Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

Answer: [\[Yes\]](#)

Justification: We have included the number and type of used GPU in the Experiment section.

Guidelines:

- The answer NA means that the paper does not include experiments.
- The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
- The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
- The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

## 9. Code Of Ethics

Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics <https://neurips.cc/public/EthicsGuidelines>?

Answer: [\[Yes\]](#)

Justification: I have read the ethics review guidelines before conducting research.

Guidelines:

- The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.
- If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.
- The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

## 10. Broader Impacts

Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

Answer: [\[Yes\]](#)

Justification: We have discussed the societal impact of our method in the appendix.

Guidelines:

- The answer NA means that there is no societal impact of the work performed.
- If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.
- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to

generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.

- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

## 11. Safeguards

Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

Answer: [NA]

Justification: This paper does not involve the use of any pretrained language models, image generators, or scraped datasets.

Guidelines:

- The answer NA means that the paper poses no such risks.
- Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
- Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
- We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

## 12. Licenses for existing assets

Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

Answer: [Yes]

Justification: we have cited the original paper that produced the code package or dataset and the used datasets in this paper are properly licensed.

Guidelines:

- The answer NA means that the paper does not use existing assets.
- The authors should cite the original paper that produced the code package or dataset.
- The authors should state which version of the asset is used and, if possible, include a URL.
- The name of the license (e.g., CC-BY 4.0) should be included for each asset.
- For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, [paperswithcode.com/datasets](https://paperswithcode.com/datasets) has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

## 13. New Assets

Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

Answer: [NA]

Justification: The datasets and models used in this paper are open-source.

Guidelines:

- The answer NA means that the paper does not release new assets.
- Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
- The paper should discuss whether and how consent was obtained from people whose asset is used.
- At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

#### 14. **Crowdsourcing and Research with Human Subjects**

Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

Answer: [NA]

Justification: The datasets are not with crowdsourcing nor research with human subjects.

Guidelines:

- The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
- Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
- According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

#### 15. **Institutional Review Board (IRB) Approvals or Equivalent for Research with Human Subjects**

Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

Answer: [NA]

Justification: The datasets are not with crowdsourcing nor research with human subjects.

Guidelines:

- The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
- Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.