Provable Discriminative Hyperspherical Embedding for Out-of-Distribution Detection

Zhipeng Zou[1,2,3], Sheng Wan[1,2,3*], Guangyu Li[1,2,3], Bo Han[4], Tongliang Liu[5], Lin Zhao[1,2,3], Chen Gong[6*]

[1]School of Computer Science and Engineering, Nanjing University of Science and Technology, China
[2]Key Laboratory of Intelligent Perception and Systems for High-Dimensional Information of Ministry of Education, China
[3]Jiangsu Key Laboratory of Image and Video Understanding for Social Security, China
[4]Hong Kong Baptist University, China
[5]Sydney AI Centre, The University of Sydney, Sydney
[6]Department of Automation, Institute of Image Processing and Pattern Recognition, Shanghai Jiao Tong University, China
chen.gong@sjtu.edu.cn, wansheng315@hotmail.com

Abstract

Out-of-distribution (OOD) detection aims to identify the test examples that do not belong to the distribution of training data. The distance-based methods, which identify OOD examples based on their distances from the centroids of in-distribution (ID) examples, have demonstrated promising OOD detection performance. However, the objectives utilized in prior approaches are typically designed for classification and thus might not yield sufficient discriminative power to distinguish between ID and OOD examples. Therefore, this paper proposes a prototype-based contrastive learning framework for OOD detection, which is termed *provable **D**iscriminative **H**yperspherical **E**mbedding (**DHE**)*. The proposed framework provides a theoretical analysis of inter-class dispersion, which is proved to be fundamental in reducing the false positive rate (FPR) on OOD examples. Based on this, we devise an angular spread loss to achieve the maximal dispersion of the prototypes of different classes prior to training. Subsequently, a prototype-enhanced contrastive loss is introduced to align embeddings of ID examples closely with their corresponding prototypes. In our proposed DHE, the maximal prototype dispersion is theoretically proved, thereby avoiding the pitfalls of local optima commonly encountered by most existing methods. Experimental results demonstrate the effectiveness of our proposed DHE, which showcases a remarkable reduction in FPR95 (*i.e.*, 5.37% on CIFAR-100) and more than doubling the computational efficiency when compared with the state-of-the-art methods.

Code — https://github.com/Canoeszzp/DHE.

Introduction

Machine learning models are typically trained with the implicit assumption that training data and test data share the same distribution, which forms in-distribution (ID) scenario. However, in many practical scenarios, a deployed neural model could be inevitably exposed to the out-of-distribution (OOD) examples that deviate from the training distribution (Rawat and Wang 2017). As a result, the model will be confused and incorrectly attribute the OOD examples into ID classes, leading to risks in practically implementing AI algorithms (Ulmer, Meijerink, and Cinà 2020; Yang et al. 2022).

To mitigate the risk caused by OOD data, OOD detection has been developed, which aims to determine whether an input example is ID or OOD. The existing OOD detection techniques can be roughly divided into four main types, *i.e.*, the confidence score-based methods (Hendrycks and Gimpel 2017; Liang, Li, and Srikant 2018; Liu et al. 2020; Zhang et al. 2022; Morteza and Li 2022), the density-based methods (Grathwohl et al. 2019; Ren et al. 2019), data augmentation-based methods (DeVries and Taylor 2017; Yun et al. 2019; Tack et al. 2020; Hendrycks et al. 2022; Wu et al. 2023; Vishwakarma, Lin, and Vinayak 2024), and the distance-based methods (Lee et al. 2018; Sehwag, Chiang, and Mittal 2020; Sun et al. 2022; Ming et al. 2023; Lu et al. 2024). Among these, the distance-based methods have shown very encouraging performance by assuming that OOD examples should be distant from the clusters of ID data in the embedding space. This assumption enables the learning of discriminative embeddings, which facilitates the accurate identification of OOD examples. Previous methods, such as SSD+ (Sehwag, Chiang, and Mittal 2020) and KNN+ (Sun et al. 2022), directly employ the existing contrastive loss (*i.e.*, SupCon) (Khosla et al. 2020) to structurize the embedding space. However, since SupCon loss is not designed specifically for OOD detection tasks, it might not yield sufficiently discriminative embeddings to distinguish between ID and OOD examples. Most recently, Ming et al. (2023) proposed a distance-based OOD detection method termed CIDER, which employs the class-conditional von Mises-Fisher (vMF) distribution (Mardia, Jupp, and Mardia 2000) to model the embeddings of ID inputs. In the training phase, CIDER utilizes a compactness loss to drive the embeddings of ID examples around their corresponding prototypes, where a dispersion loss is employed to ensure separation among different prototypes.

*Corresponding authors: Chen Gong, Sheng Wan.

Previous works (Sehwag, Chiang, and Mittal 2020; Sun et al. 2022; Ming et al. 2023) have demonstrated that large inter-class dispersion helps to improve the performance of OOD detection. Nevertheless, current distance-based OOD detection methods did not adequately explore the theoretical foundations for the effectiveness of prototype dispersion. As a result, the existing distance-based methods could be trapped in local optima when conducting inter-class dispersion, which leads to performance degradation and considerable waste of computing resources. Therefore, in this work, we propose a simple yet effective distance-based OOD detection method called provable **D**iscriminative **H**yperspherical **E**mbedding (**DHE**) to obtain the embeddings that are highly discriminative in distinguishing ID and OOD examples. Specifically, we conduct an in-depth theoretical analysis of inter-class dispersion, which demonstrates that increasing inter-class dispersion is beneficial for reducing the false positive rate (FPR) of model on OOD examples. Inspired by these theoretical insights, we introduce an angular spread loss to maximize prototype dispersion. Additionally, a prototype-enhanced contrastive (PEC) loss is utilized to ensure that the embeddings of ID examples are closely around their corresponding prototypes, which further enhances the discriminability of feature embeddings. By this means, the proposed method theoretically guarantees the maximization of prototype dispersion, which leads to a more reliable model than the previous models without theoretical foundations. Besides, since the prototypes with maximal dispersion are efficiently pre-computed before iterative classifier training, the computation burden of our method is significantly reduced when compared with existing methods.

It is worth noting that although our proposed DHE looks similar to CIDER (Ming et al. 2023), they diverges fundamentally in multiple key aspects. Specifically, our DHE theoretically ensures the maximization of dispersion among different class prototypes. In contrast, CIDER cannot guarantee such maximal inter-class dispersion among prototypes. Additionally, we theoretically prove that increasing inter-class distance can enhance the ability to detect OOD examples when a distance-based scoring function is adopted. However, such theoretical justification is absent in CIDER. Consequently, when compared with CIDER, our DHE achieves enhanced training efficiency and superior OOD detection performance. The contributions of this paper are summarized as follows:

- We provide new insights for distance-based OOD detection methods, which theoretically reveal that the inter-class dispersion enhancement is helpful for improving the OOD detection performance.
- We propose a simple yet effective prototype-based contrastive learning framework termed provable **D**iscriminative **H**yperspherical **E**mbedding (**DHE**), which can theoretically guarantee the maximization of inter-class prototype dispersion. Note that the prototypes of our method can be efficiently pre-computed without any complicated and time-consuming network optimization.
- Extensive experiments demonstrate the superiority of the proposed DHE over existing methods in terms of both false positive fate and efficiency. For example, CIFAR-100 (Krizhevsky, Hinton et al. 2009) dataset, our method surpasses the state-of-the-art method (*i.e.*, CIDER), by 5.37% in FPR95 and only needs approximately half of the computational time of CIDER.

Theoretical Implication

We consider multi-class classification, where $\mathcal{X}$ denotes the input space and $\mathcal{Y}_{in} = \{1, 2, ..., K\}$ denotes the label space of ID data, with K denoting the total number of categories in the training data. We assume access to the labeled training set $\mathcal{D}_{tr} = \{(\mathbf{x}_i, y_i)\}_{i=1}^{N}$, where $\mathbf{x}_i \in \mathcal{X}$ and $y_i \in \mathcal{Y}_{in}$ are drawn *i.i.d.* from the joint distribution $\mathcal{P}_{\mathcal{X}\times\mathcal{Y}_{in}}$. Here, N is the size of training set. We also denote $\mathcal{P}_{in}$ as the marginal distribution on $\mathcal{X}$.

In open-world scenarios, machine learning models often encounter OOD examples with labels y_{out} that are not present in the training data. That is to say, we have $y_{out} \notin \mathcal{Y}_{in}$, which indicates that there is no overlap between the label space of ID and OOD data. In other words, the label space of OOD data, denoted as $\mathcal{Y}_{out}$ does not intersect with the label space of ID data *i.e.*, $\mathcal{Y}_{in} \cap \mathcal{Y}_{out} = \emptyset$. The aim of OOD detection is to identify whether an example $\mathbf{x} \in \mathcal{X}$ is from $\mathcal{P}_{in}$ (ID) or not (OOD). The decision can be made via a level set estimation:

$$G_\tau(\mathbf{x}) = \begin{cases} \text{ID data} & S_\theta(\mathbf{x}) \geq \tau \\ \text{OOD data} & S_\theta(\mathbf{x}) < \tau \end{cases}, \tag{1}$$

where $S_\theta(\cdot)$ is the scoring function related to the neural network parameter θ, and τ represents the threshold. In Eq. (1), examples are determined as ID data if their scores $S_\theta(\mathbf{x}) \geq \tau$ and as OOD data, otherwise. Here, setting a reasonable value for τ can typically contribute to high identification accuracy.

The challenges of OOD detection encompass two aspects, namely: 1) preventing ID examples from being erroneously judged as OOD examples by the level set estimation (*i.e.*, Eq. (1)), and 2) correctly identifying input OOD data with a high probability. Therefore, to address the above challenges and provide guidance for the subsequent algorithm design, we conduct some useful theoretical analyses here.

To cope with the first challenge, we first establish a formal definition of the threshold set $\mathcal{T}$ as

Definition 1. *We denote the set of thresholds as* $\mathcal{T} = \{\tau : P\left(S_\theta(\mathbf{x}_i) < \tau\right) \leq \alpha,\ \mathbf{x}_i \in \mathcal{P}_{in}\}$*, where* $P(\cdot)$ *denotes probability throughout this paper. These thresholds ensure that the probability of mis-identifying an ID example* $\mathbf{x}_i$ *from* $\mathcal{P}_{in}$ *as OOD is less than a specified probability* α *(e.g., 0.05).*

To address the first challenge, *i.e.*, preventing ID examples from being judged as OOD, it is essential to select an appropriate threshold τ in Definition 1. For a given probability α, a small threshold is preferred, as it provides a tight level set estimation that can decrease the probability of misidentifying an ID example as OOD.

Lemma 2. *Given a small probability* α*, an example* $\mathbf{x}_i \in \mathcal{D}_{tr}$*,* $\tau \in \mathcal{T}$*, and the scoring function* $S_\theta(\cdot)$*, then we have*

$$\tau \leq \mathbb{E}\left(S_\theta(\mathbf{x}_i)\right) - \sigma\left(S_\theta(\mathbf{x}_i)\right)/\sqrt{\alpha}, \tag{2}$$

where $\mathbb{E}(\cdot)$ *represents the mathematical expectation, and* $\sigma(\cdot)$ *is standard deviation. The equality holds if and*

only if $P\left(S_\theta(\mathbf{x}_i) > 2\mathbb{E}\left(S_\theta(\mathbf{x}_i)\right) - \tau_\alpha\right) = 0$, *where* $\tau_\alpha = \mathbb{E}\left(S_\theta(\mathbf{x}_i)\right) - \sigma\left(S_\theta(\mathbf{x}_i)\right)/\sqrt{\alpha}$ *is the maximum threshold for a given probability* α.

The proof of the Lemma 2 is presented in Appendix A.1. Based on Lemma 2, we can derive the level set estimation with the threshold $\tau \leq \tau_\alpha$, which guarantees that the probability of mis-identifying an ID example as OOD is lower than α. Therefore, the first challenge of OOD detection can be addressed.

Building upon Lemma 2, we then proceed to address the second challenge in OOD detection. The goal here is to control the error rate of an OOD example being mis-identified as an ID one, and this is commonly evaluated by the FPR metric. To achieve this, we aim to reveal the relationship of FPR with inter-class distance, and that with intra-class distance. Formally, we denote $f_\theta : \mathcal{X} \to \mathcal{R}^D$ is the encoder with θ representing the network parameters and D being the embedding dimension, respectively. The embedding of input example is denoted as $\mathbf{z}_i = f_\theta(\mathbf{x}_i)$. For a given class $c \in \{1, 2, ..., K\}$, the class prototype is denoted as $\boldsymbol{\mu}_c = \mathbb{E}(\mathbf{z}_i | y_i = c)$.

Theorem 3. *If the scoring function* $S_\theta : \mathcal{R}^D \to \mathcal{R}$ *is distance-based, then the FPR for estimating* $P(G_{\tau_\alpha}(\mathbf{x}_0) =$ *ID data*$)$ *has* $FPR \propto \hat{r}_n/\hat{r}_o$, *where* $\mathbf{x}_0$ *refers to an OOD example,* $\hat{r}_n = \mathbb{E}\left(\mathbb{E}\left(\|\mathbf{z}_i - \boldsymbol{\mu}_c\|_2 \,|\, y_i = c\right)\right)$ *denotes the average intra-class distances, and* $\hat{r}_o = \mathbb{E}_{c_1 \neq c_2}\left(\|\boldsymbol{\mu}_{c_1} - \boldsymbol{\mu}_{c_2}\|_2\right), c_1, c_2 \in \{1, 2, ..., K\}$ *represents the average inter-class distances.*

Theorem 3 indicates that the FPR is proportional to $\hat{r}_n$ and is inversely proportional to $\hat{r}_o$ when using a distance-based scoring function. To reduce the FPR in OOD detection, we can decrease $\hat{r}_n$ and increase $\hat{r}_o$. Given that no OOD data are involved during the training phase, Theorem 3 offers strong theoretical guidance on constructing discriminative embeddings $\mathbf{z}$ for the data $\mathbf{x}$. The proof of Theorem 3 is presented in Appendix A.2.

Method

Building on the insights from Theorem 3, we propose a training framework for acquiring suitable data embedding, so that the second challenge of OOD detection can be addressed from two crucial aspects, namely: 1) maximizing inter-class distances to enhance category distinction; and 2) ensuring the feature embeddings are closely around the corresponding prototypes of the same class. The framework of our method is shown in Figure 1. Firstly, prior to classifier training, we initialize the prototypes by averaging the embeddings $\mathbf{z}_i$ of each class. Subsequently, we optimize the dispersion among prototypes to obtain a set of maximally dispersed prototypes $\mathcal{M} = \{\boldsymbol{\mu}_c \in \mathcal{R}^D, c \in \{1, 2, ..., K\}\}$. During the classifier training phase, the encoder f_θ is trained to ensure that the embeddings of ID examples are closely around their corresponding prototypes.

Construction of Hyperspherical Embedding

We establish the embeddings using a hyperspherical model, inspired by the benefits highlighted in (Khosla et al. 2020; Wang and Isola 2020). The embedding $\mathbf{z}$ is positioned on a unit hypersphere ($\|\mathbf{z}_i\|_2 = 1$) and is modeled via the von Mises-Fisher (vMF) distribution (Mardia, Jupp, and Mardia 2000). Here, the probability density function of $\mathbf{z}$ is positioned on the hypersphere can be defined as

$$p_D(\mathbf{z}_i; \boldsymbol{\mu}_c, \kappa) = Z_d(\kappa) \exp\left(\kappa \mathbf{z}_i^\top \boldsymbol{\mu}_c\right), \tag{3}$$

where $\kappa \geq 0$ measures the concentration of the embeddings around the prototype $\boldsymbol{\mu}_c$, and $Z_d(\kappa)$ is a normalization factor. As κ increases, the distribution of embeddings becomes more concentrated around the corresponding $\boldsymbol{\mu}_c$. When κ approaches to 0, the embeddings are uniformly distributed across the hypersphere. Based on this, the normalized probability that the embedding $\mathbf{z}_i$ belongs to category c can be expressed as

$$\begin{aligned} P(y_i = c \,|\, \mathbf{z}_i; \{\kappa, \boldsymbol{\mu}_j\}_{j=1}^K) &= \frac{Z_d\left(\kappa\right) \exp\left(\kappa \mathbf{z}_i^\top \boldsymbol{\mu}_c\right)}{\sum_{j=1}^K Z_d(\kappa) \exp\left(\kappa \mathbf{z}_i^\top \boldsymbol{\mu}_j\right)} \\ &= \frac{\exp\left(\mathbf{z}_i^\top \boldsymbol{\mu}_c / t\right)}{\sum_{j=1}^K \exp\left(\mathbf{z}_i^\top \boldsymbol{\mu}_j / t\right)}, \end{aligned} \tag{4}$$

where $t = 1/\kappa$ acts similarly to a temperature parameter.

Prototype Dispersion Maximization

To achieve the first training objective detailed at the beginning of this section, namely maximizing inter-class distances to enhance category distinction, we focus on the optimization of class prototypes and explore the conditions that maximize inter-class distance. To this end, we first provide the following Theorem 4

Theorem 4. *For any two classes* $i, j \in \{1, 2, ..., K\}$, *the sum of squared distances between prototypes* $\boldsymbol{\mu}_i$ *and* $\boldsymbol{\mu}_j$ *is upper bounded by*

$$\frac{1}{2} \sum\nolimits_{i \neq j} \|\boldsymbol{\mu}_i - \boldsymbol{\mu}_j\|_2^2 \leq K^2, \tag{5}$$

where the equality holds if and only if

$$\boldsymbol{\mu}_i^\top \boldsymbol{\mu}_j = \begin{cases} 1, & i = j \\ 1/(1-K), & i \neq j \end{cases}. \tag{6}$$

Theorem 4 specifies the optimal conditions for maximizing the distance between different class prototypes, and its proof is available in Appendix A.3.

Motivate by Theorem 4, we introduce an angular spread loss to encourage maximal dispersion among class prototypes,which is

$$\mathcal{L}_{\text{as}} = \frac{1}{K} \sum_{i=1}^K \log \frac{1}{K-1} \sum_{j=1}^K \mathbb{I}\{i \neq j\} e^{\left[\boldsymbol{\mu}_i^\top \boldsymbol{\mu}_j - 1/(1-K)\right]^2 / t}, \tag{7}$$

where t is a temperature parameter. Here, the indicator function $\mathbb{I}\{\cdot\}$ equals 1 if the argument inside the bracket holds, and 0 otherwise. Minimizing the angular spread loss $\mathcal{L}_{\text{as}}$ is equivalent to finding the global optimum of a quadratic function, which ensures that the maximal dispersion of prototypes can be achieved efficiently and reliably. Therefore, the maximal dispersion of class prototypes can be guaranteed

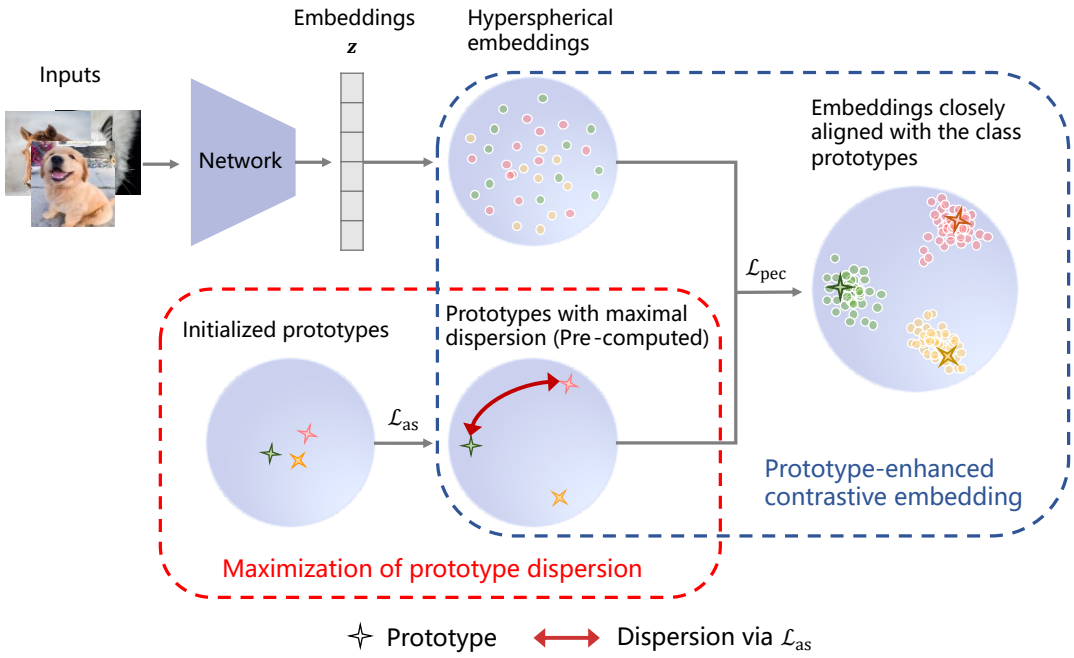

Figure 1: Overview of our proposed framework. Prior to the training phase, the dispersion between class prototypes is maximized by optimizing the angular spread loss. During model training, we further minimize our prototype-enhanced contrastive (PEC) loss to encourage the embedding of ID examples to align closely with their corresponding class prototypes. As a result, discriminative hyperspherical embeddings can be obtained to enhance the distinction between ID examples and OOD examples.

both theoretically and empirically. In practice, we firstly use $\mathcal{D}_{tr} = \{(\mathbf{x}_i, y_i)\}_{i=1}^N$ to initialize the prototypes, and then calculate the loss $\mathcal{L}_{\text{as}}$. Afterward, the prototypes will be updated based on the gradient of the $\mathcal{L}_{\text{as}}$, in order to maximize the distance among the class prototypes. The specific optimization process has been explained in Appendix B. Note that prototype initialization and optimization are performed prior to the subsequent classifier training, and the prototypes are kept unchanged throughout the training of classifier.

Prototype-Enhanced Contrastive Embedding

To achieve the second training objective, *i.e.*, ensuring the embeddings of ID examples are closely around their corresponding prototypes sharing the same class, we use the training dataset $\mathcal{D}_{tr} = \{(\mathbf{x}_i, y_i)\}_{i=1}^N$ to perform maximum likelihood estimation (MLE), which is formulated as

$$\arg\max_{\theta} \prod_{i=1}^{N} p(y_i|\mathbf{z}_i; \{\kappa, \boldsymbol{\mu}_c\}_{c=1}^K), \tag{8}$$

where $\mathbf{z}_i$ is the embedding of $\mathbf{x}_i$, and $\boldsymbol{\mu}_c$ belongs to the set of class prototypes $\mathcal{M} = \{\boldsymbol{\mu}_c, c \in \{1, 2, ..., K\}\}$. Thanks to the universal approximation power of neural networks (Hornik, Stinchcombe, and White 1989), we propose a prototype-based contrastive learning method for solving the MLE problem in Eq. (8). Specifically, we introduce a prototype-enhanced contrastive (PEC) loss to encourage embeddings to closely align with their class prototypes, which can be expressed as

$$\mathcal{L}_{\text{pec}} = -\frac{1}{N} \sum_{i=1}^{N} \sum_{c=1}^{K} \mathbb{I}\{y_i = c\} \log\left(p_i^c\right). \tag{9}$$

Here, p_i^c quantifies the normalized proximity-based probability between the embedding $\mathbf{z}_i$ and the corresponding class prototype $\boldsymbol{\mu}_c$, which is denoted as

$$p_i^c = \frac{\exp\left(\mathbf{z}_i^\top \boldsymbol{\mu}_c / t\right)}{\sum_{j=1}^{K} \exp\left(\mathbf{z}_i^\top \boldsymbol{\mu}_j / t\right)}.$$

The employment of PEC loss pushes the data embeddings of the same class close to their corresponding prototypes, which satisfies the second training objective mentioned above.

To summarize, the proposed method theoretically guarantees the maximization of prototype dispersion through the optimization of the proposed angular spread loss $\mathcal{L}_{\text{as}}$. Subsequently, the utilization of PEC loss $\mathcal{L}_{\text{pec}}$ helps align the embeddings of input examples with their corresponding class prototypes, which enhances the compactness of the intra-class embeddings. In a word, the effectiveness of our method in OOD detection can be primarily attributed to the theoretical guarantee for maximal prototype dispersion and the tight clustering of inter-class embeddings. Further details about the entire training framework are provided in Appendix B.

Experiments

In this section, we present a series of experiments designed to demonstrate the effectiveness of the proposed DHE in OOD detection. We compare DHE with multiple state-of-the-art methods across various benchmarks, including CIFAR-10 (Krizhevsky, Hinton et al. 2009), CIFAR-100 (Krizhevsky, Hinton et al. 2009), and ImageNet-100 (Deng et al. 2009). Our experiments are designed to validate the theoretical

advantages of DHE, especially its ability to enhance discriminability between ID and OOD examples. Additionally, we evaluate the computational efficiency to verify that our method can achieve reduced training time without compromising performance.

Experimental Setup

Datasets and training details. We use the CIFAR-10 (Krizhevsky, Hinton et al. 2009) and CIFAR-100 (Krizhevsky, Hinton et al. 2009) as our ID datasets, which have been commonly adopted in this field. For evaluation of OOD detection, we use five commonly-used datasets, including SVHN (Netzer et al. 2011), Places365 (Zhou et al. 2017), Texture (Cimpoi et al. 2014), LSUN (Yu et al. 2015), and iSUN (Xu et al. 2015). In our main experiments, ResNet-18 is employed as the backbone for CIFAR-10, and ResNet-34 is deployed on CIFAR-100. The model is trained by using stochastic gradient descent (SGD) with the momentum and the weight decay setting to 0.9 and 10^{-4}, respectively. Besides, we keep our hyperparameters the same as those used in CIDER (Ming et al. 2023) and ReweightOOD (Regmi et al. 2024b). Specifically, we set the initial learning rate to 0.5 with cosine scheduling, maintain a batch size of 512, and conduct training for a duration of 500 epochs. The embedding dimension D is set to 128 for our projector, which is also consistent with the existing research (Khosla et al. 2020; Sun et al. 2022; Ming et al. 2023). The temperature t in our method is set to 0.1. Additional experimental details are provided in Appendix C.1.

OOD detection score. Our framework is designed to learn discriminative representations. In our main experiments, we evaluate the performance of our method using KNN (Sun et al. 2022), which is a non-parametric distance-based OOD detection approach. Concretely, if the distance between the input example and its K-th nearest example in the training set exceeds a predetermined threshold, the example is classified as OOD. Since the features of all examples are normalized to the unit norm, the distance metric here becomes the cosine similarity between feature vectors. To ensure fairness in comparison, we also employ the widely used Mahalanobis distance (Lee et al. 2018) for OOD data judgement.

Evaluation metric. To reveal the effectiveness of the proposed DHE, we utilize two common metrics: 1) the false positive rate of OOD examples when the true positive rate of ID examples is at 95% (*i.e.*,FPR95), and 2) the area under the receiver operating characteristic curve (AUROC).

Main Results

DHE outperforms different baseline methods. In Table 1, we present the outcomes of our experiments conducted under the standard setting, where CIFAR-100 serves as the ID dataset and other datasets are deemed as OOD data. To ensure a fair comparison, we employ ResNet-34 trained on the CIFAR-100 (ID) dataset, without accessing to any other external OOD datasets. We compare the proposed DHE with two categories of methods: post-hoc methods and training methods. The post-hoc methods include MSP (Hendrycks and Gimpel 2017), ODIN (Liang, Li, and Srikant 2018), Mahalanobis (Lee et al. 2018), and Energy (Liu et al. 2020). The training methods include GODIN (Hsu et al. 2020), T2FNORM (Regmi et al. 2024a), and several related contrastive learning methods such as SimCLR (Chen et al. 2020), CSI (Tack et al. 2020), SSD+ (Sehwag, Chiang, and Mittal 2020), KNN+ (Sun et al. 2022), CIDER (Ming et al. 2023), and ReweightOOD (Regmi et al. 2024b).

Type	Method	Average (%)	
		FPR↓	AUROC↑
Post-hoc	MSP (ICLR'17)	83.57	75.27
	ODIN (ICLR'18)	78.70	78.91
	Mahalanobis (NeurIPS'18)	80.15	79.53
	Energy (NeurIPS'20)	70.72	82.55
Training	GODIN (CVPR'20)	87.57	70.97
	CE+SimCLR (ICML'20)	59.62	84.15
	CSI (NeurIPS'20)	67.48	84.83
	SSD+ (ICLR'20)	62.33	86.64
	KNN+ (ICML'22)	62.21	86.39
	CIDER (ICLR'23)	48.89	87.39
	T2FNORM (CVPR'24)	69.07	83.01
	ReweightOOD (CVPR'24)	56.74	86.03
	DHE (ours)	**43.52**	**87.82**

Table 1: Comparison of OOD detection performance averaged over five OOD benchmarks when CIFAR-100 is adopted as ID dataset. "↑" denotes larger values are better, and "↓" indicates smaller values are better. **Bold** numbers indicate the best results. The specific result on each of the five datasets is displayed in Appendix C.2.

As shown in Table 1, the proposed DHE significantly enhances the OOD detection performance and achieves superior performance than the baseline methods. Unlike the existing distance-based approaches that employ contrastive loss such as KNN+ and SSD+, DHE can effectively maximize the inter-class dispersion specifically for OOD detection. As a result, DHE achieves a reduction of 18.81% compared with SSD+ and 18.69% compared with KNN+ on FPR95, respectively. Moreover, DHE outperforms the latest baseline methods T2FNORM and ReweightOOD, reducing FPR95 by 25.55% and 13.22%, respectively. Besides, DHE surpasses the most relevant baseline method, *i.e.*, CIDER, by 5.37% on FPR95. Note that the main difference between the proposed DHE and CIDER lies in that DHE is provable to obtain the prototypes with maximal dispersion, which can greatly enhance the discriminative power of embeddings. Although CIDER attempts to optimize inter-class prototype dispersion during the training of classifier, it is achieved by an iterative optimization process and may be trapped in local optima. More experimental results when CIFAR-10 as ID data are presented in Appendix C.2.

DHE demonstrates effectiveness across different distance-based scores. The comparison of different OOD detection scores is exhibited in Table 2. Here, we consider two commonly used scoring functions in OOD detection, namely KNN (Sun et al. 2022) and Mahalanobis distance (Lee et al. 2018). Under both KNN (non-parametric) and Mahalanobis

Metric	Method	Average (%)	
		FPR↓	AUROC↑
KNN	KNN+	62.21	86.39
	CIDER	48.89	87.39
	ReweightOOD	56.74	86.03
	DHE(ours)	**43.52**	**87.82**
Mahalanobis	SSD+	62.33	86.64
	CIDER	49.37	87.98
	ReweightOOD	53.94	88.25
	DHE(ours)	**44.61**	**89.01**

Table 2: Results obtained by adopting different distance-based scores when using CIFAR-100 as ID dataset. The results averaged over five OOD benchmarks are reported. "↑" denotes larger values are better and "↓" indicates smaller values are better. **Bold** numbers indicate the best results.

distance (parametric) scores, DHE outperforms the existing approaches. Specifically, when adopting KNN distance, DHE outperforms KNN+ (SupCon + KNN) by 18.69%, CIDER by 5.37%, and ReweightOOD by 13.22% in terms of FPR95, respectively. Additionally, when employing Mahalanobis distance, DHE surpasses SSD+ (SupCon + Mahalanobis) by 18.81%, CIDER by 4.76%, and ReweightOOD by 9.33% in terms of FPR95, respectively. All these statistics indicate the effectiveness of the proposed DHE under different distance-based scores. The detailed results on each dataset related to Table 2 are presented in Appendix C.2.

DHE is competitive on large-scale datasets. To evaluate the performance of DHE in more realistic scenarios, experiments are performed on challenging large-scale benchmarks. Specifically, we used ImageNet-100 as the ID dataset, which is a subset of ImageNet (Deng et al. 2009) consisting of 100 randomly sampled classes. Meanwhile, we employed the same OOD datasets as those adopted in CIDER (Ming et al. 2023), including subsets of iNaturalist (Van Horn et al. 2018), SUN (Xiao et al. 2010), Places365 (Zhou et al. 2017), and Texture (Cimpoi et al. 2014). To improve the efficiency, we fine-tuned a pre-trained ResNet-50 for 10 epochs with an initial learning rate of 0.005. Concretely, we focus on the parameters of the last residual block and the projector, while freezing the weights in other modules. The performance (in AUROC) is shown in Figure 2, where it can be observed that DHE outperforms CIDER and other distance-based methods across all datasets. This further demonstrates the advantage of our approach in maximizing prototype dispersion. Detailed experimental results are available in Appendix C.2.

Discussions

DHE enhances the discrimination between ID and OOD examples. We visualize the embedding distributions of ID (CIFAR-10) and OOD (LSUN) data using UMAP (McInnes et al. 2018) in Figure 3. A notable observation is that the embeddings obtained by DHE exhibit better discriminability between ID and OOD embeddings when compared with those trained with cross-entropy (CE) loss. Additionally, we estimated the density distribution of ID and OOD examples

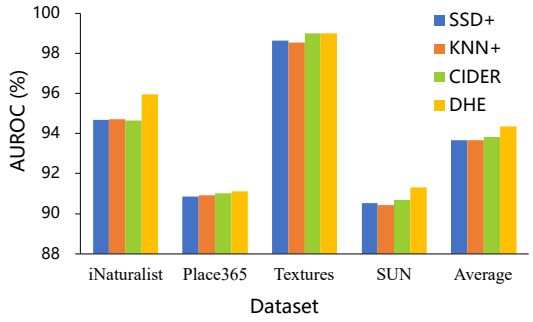

Figure 2: OOD detection performance obtained by fine-tuning the pre-trained ResNet-50 model on ImageNet-100.

regarding KNN scores in Figure 3, which further verifies the strong performance of DHE in separating ID and OOD data.

DHE exhibits high computational efficiency. Figure 4 exhibits the training time (seconds per epoch) of different methods when using the CIFAR-100 dataset (ID) with a ResNet-34 model. Here, we compare our proposed DHE with the CE loss and the popular contrastive learning method SupCon (Khosla et al. 2020) that is utilized in KNN+ (Sun et al. 2022) and SSD+ (Sehwag, Chiang, and Mittal 2020). We also compare our proposed DHE with CIDER (Ming et al. 2023), which is specially designed for OOD detection. The results clearly show that DHE maintains competitive computational efficiency, when compared with CE and SupCon. It is worth noting that our method reduces training time by 62% when compared with CIDER. This is due to that our method achieves maximal prototype dispersion before training, while CIDER continuously updates prototypes using an exponential-moving-average (EMA) approach (Grathwohl et al. 2020) during training. This ongoing optimization in CIDER introduces significant computational overhead.

Related Work

Out-of-Distribution Detection

In recent years, with the flourishing development of the machine learning community, OOD detection has attracted increasing attention. The key to OOD detection is the development of an effective ID-OOD binary classifier (Yang et al. 2021; Szyc, Walkowiak, and Maciejewski 2023). The foundation of OOD detection research stems from the approach that uses the maximum softmax probability (Hendrycks and Gimpel 2017) to identify OOD examples. Subsequently, various methods for OOD detection have emerged, such as the gradient-based methods (Liang, Li, and Srikant 2018; Huang, Geng, and Li 2021) and density-based methods (Grathwohl et al. 2019; Ren et al. 2019). In the meanwhile, methods deriving improved OOD scores based on neural network outputs (Hendrycks and Gimpel 2017; Liang, Li, and Srikant 2018; Hsu et al. 2020; Liu et al. 2020; Wang et al. 2021; Zhang et al. 2022; Regmi et al. 2024a) and utilizing strong data augmentation (DeVries and Taylor 2017; Yun et al. 2019; Hendrycks et al. 2019; Thulasidasan et al. 2019; Mohseni et al. 2020; Tack et al. 2020; Ahmadian, Lindsten, and Zhou 2021; Hendrycks et al. 2022; Wang et al. 2022; Wu et al.

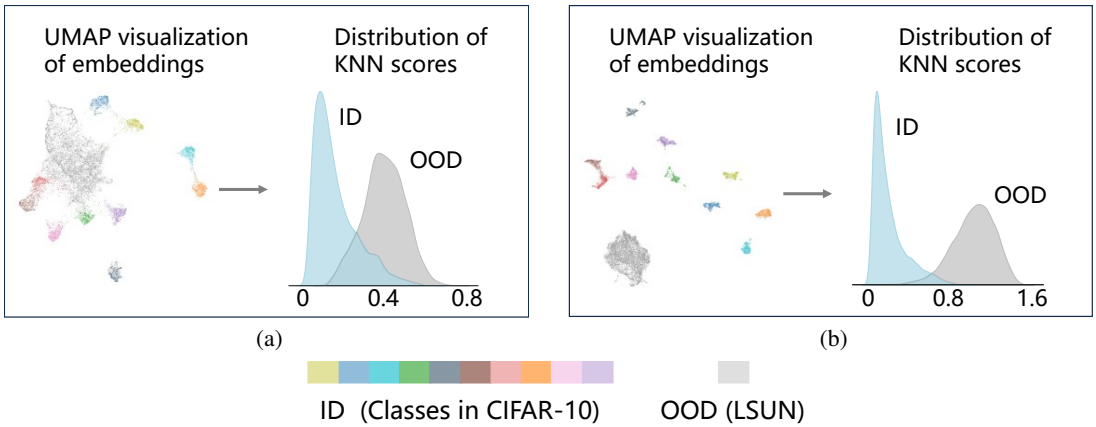

Figure 3: Visualization of embeddings from the penultimate layer using UMAP (McInnes et al. 2018). The models are trained with ResNet-18 using (a) CE loss and (b) our proposed DHE method. The scoring function is KNN. The ID dataset is CIFAR-10 with ID classes depicted in various colors, and the OOD dataset is LSUN with OOD data depicted in gray.

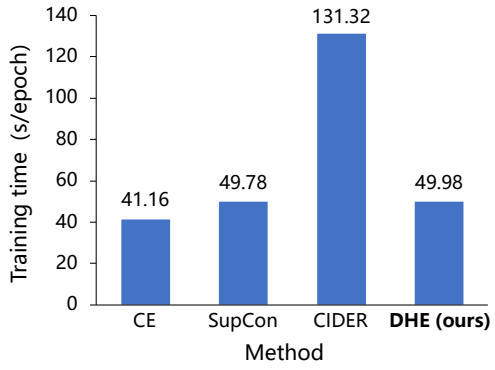

Figure 4: Training time of different methods.

2023; Vishwakarma, Lin, and Vinayak 2024) have also shown promising results. Among these newly developed OOD detection techniques, the distance-based methods (Lee et al. 2018; Sehwag, Chiang, and Mittal 2020; Tack et al. 2020; Ren et al. 2021; Sun et al. 2022; Ming et al. 2023; Ghosal, Sun, and Li 2024) have gained intensive attention due to their theoretical clarity and effectiveness. They are established based on the principle that OOD examples should demonstrate substantial separation from the centroids or prototypes of ID classes during testing. Our work contributes to this area by designing a distance-based OOD detection approach within a hyperspherical embedding space.

Contrastive Learning

Contrastive learning is a self-supervised learning method that aims to learn representations by comparing the similarities and differences between data examples. It can be leveraged to enhance OOD detection by training models to maximize the similarity within the same class and minimize the similarity across different classes, simultaneously. Recent advancements in contrastive representation learning methods, such as SimCLR (Chen et al. 2020) and SupCon (Khosla et al. 2020), have paved the way for distance-based approaches for OOD detection. For example, the CSI (Tack et al. 2020) examines the impact of various data augmentations on OOD detection via SimCLR. Besides, SSD+ (Sehwag, Chiang, and Mittal 2020) and KNN+ (Sun et al. 2022) leverage SupCon to construct embeddings that are more effective for OOD detection. Furthermore, methods such as VOS (Du et al. 2022b) and NPOS (Tao et al. 2023) enhance OOD detection by contrasting synthesized OOD examples with training data to refine the decision boundaries between ID and OOD examples. Additionally, some recent works (Du et al. 2022a; Ming et al. 2023; Tao et al. 2023; Lu et al. 2024) that adopt vMF distribution (Mardia, Jupp, and Mardia 2000) for data modeling provide clear insights into hyperspherical embedding. Specifically, CIDER (Ming et al. 2023) proposed an optimization strategy that pushes examples from the same class close to their corresponding prototypes while ensuring maximal dispersion among different classes. However, minimizing the dispersion loss may lead to convergence to local minima and requires extensive computational resources. To address this challenge, our proposed method theoretically ensures the attainment of globally optimal dispersed prototypes before the classifier training. This pre-training strategy not only enhances the OOD detection performance but also significantly reduces computational demands.

Conclusion

In this work, we propose DHE, a simple yet effective prototype-based contrastive learning framework for OOD detection. Our theoretical analysis demonstrates that inter-class dispersion is crucial for effectively distinguishing between ID and OOD examples. Inspired by this, we devise an angular spread loss to provably maximize the dispersion among prototypes. Furthermore, we introduce a prototype-enhanced contrastive loss to ensure that embeddings are tightly clustered around their corresponding class prototypes. By simultaneously maximizing the inter-class distances and minimizing the intra-class distance, the ID-OOD separability can be greatly enhanced. Building on the above-mentioned theoretical foundation, our empirical evaluations reveal that DHE exhibits superior OOD detection performance and computational efficiency on common OOD benchmarks, when compared with the state-of-the-art baseline methods.

Acknowledgments

This research is supported by the National Natural Science Foundation of China (NSFC) under Grants Nos. 62336003, 12371510, 62172222, and 62006119; the NSF for Distinguished Young Scholars of Jiangsu Province (No. BK20220080); the NSF of Jiangsu Province (No. BK20241469); the Postdoctoral Fellowship Program of the China Postdoctoral Science Foundation (CPSF) (No. GZC20233503); the Project funded by the China Postdoctoral Science Foundation (Nos. 2023M741708, 2023TQ0159); and the National Key Research and Development Program of China (International Collaboration Special Project, No. SQ2023YFE0102775).

References

Ahmadian, A.; Lindsten, F.; and Zhou, Z.-H. 2021. Likelihood-Free Out-of-Distribution Detection With Invertible Generative Models. In *IJCAI*, 2119–2125.

Chen, T.; Kornblith, S.; Norouzi, M.; and Hinton, G. 2020. A Simple Framework for Contrastive Learning of Visual Representations. In *ICML*, 1597–1607. PMLR.

Cimpoi, M.; Maji, S.; Kokkinos, I.; Mohamed, S.; and Vedaldi, A. 2014. Describing Textures in the Wild. In *CVPR*, 3606–3613.

Deng, J.; Dong, W.; Socher, R.; Li, L.-J.; Li, K.; and Fei-Fei, L. 2009. Imagenet: A Large-Scale Hierarchical Image Database. In *CVPR*, 248–255. Ieee.

DeVries, T.; and Taylor, G. W. 2017. Improved Regularization of Convolutional Neural Networks With Cutout. *arXiv preprint arXiv:1708.04552*.

Du, X.; Gozum, G.; Ming, Y.; and Li, Y. 2022a. Siren: Shaping Representations for Detecting Out-of-Distribution Objects. *NeurIPS*, 35: 20434–20449.

Du, X.; Wang, Z.; Cai, M.; and Li, Y. 2022b. VOS: Learning What You Don't Know by Virtual Outlier Synthesis. In *ICLR*.

Ghosal, S. S.; Sun, Y.; and Li, Y. 2024. How to Overcome Curse-of-Dimensionality for Out-of-Distribution Detection? In *AAAI*, volume 38, 19849–19857.

Grathwohl, W.; Wang, K.-C.; Jacobsen, J.-H.; Duvenaud, D.; Norouzi, M.; and Swersky, K. 2019. Your Classifier Is Secretly an Energy Based Model and You Should Treat It Like One. In *ICLR*.

Grathwohl, W.; Wang, K.-C.; Jacobsen, J.-H.; Duvenaud, D.; Norouzi, M.; and Swersky, K. 2020. Mopro: Webly Supervised Learning With Momentum Prototypes. In *ICLR*.

Hendrycks, D.; and Gimpel, K. 2017. A Baseline for Detecting Misclassified and Out-of-Distribution Examples in Neural Networks. In *ICLR*.

Hendrycks, D.; Mu, N.; Cubuk, E. D.; Zoph, B.; Gilmer, J.; and Lakshminarayanan, B. 2019. Augmix: A Simple Data Processing Method to Improve Robustness and Uncertainty. In *ICLR*.

Hendrycks, D.; Zou, A.; Mazeika, M.; Tang, L.; Li, B.; Song, D.; and Steinhardt, J. 2022. Pixmix: Dreamlike Pictures Comprehensively Improve Safety Measures. In *CVPR*, 16783–16792.

Hornik, K.; Stinchcombe, M.; and White, H. 1989. Multilayer Feedforward Networks Are Universal Approximators. *Neural Networks*, 2(5): 359–366.

Hsu, Y.-C.; Shen, Y.; Jin, H.; and Kira, Z. 2020. Generalized Odin: Detecting Out-of-Distribution Image Without Learning From Out-of-Distribution Data. In *CVPR*, 10951–10960.

Huang, R.; Geng, A.; and Li, Y. 2021. On the Importance of Gradients for Detecting Distributional Shifts in the Wild. *NeurIPS*, 34: 677–689.

Khosla, P.; Teterwak, P.; Wang, C.; Sarna, A.; Tian, Y.; Isola, P.; Maschinot, A.; Liu, C.; and Krishnan, D. 2020. Supervised Contrastive Learning. *NeurIPS*, 33: 18661–18673.

Krizhevsky, A.; Hinton, G.; et al. 2009. Learning Multiple Layers of Features From Tiny Images.

Lee, K.; Lee, K.; Lee, H.; and Shin, J. 2018. A Simple Unified Framework for Detecting Out-of-Distribution Samples and Adversarial Attacks. *NeurIPS*, 31.

Liang, S.; Li, Y.; and Srikant, R. 2018. Enhancing the Reliability of Out-of-Distribution Image Detection in Neural Networks. In *ICLR*.

Liu, W.; Wang, X.; Owens, J.; and Li, Y. 2020. Energy-Based Out-of-Distribution Detection. *NeurIPS*, 33: 21464–21475.

Lu, H.; Gong, D.; Wang, S.; Xue, J.; Yao, L.; and Moore, K. 2024. Learning With Mixture of Prototypes for Out-of-Distribution Detection. In *ICLR*.

Mardia, K. V.; Jupp, P. E.; and Mardia, K. 2000. *Directional Statistics*, volume 2. Wiley Online Library.

McInnes, L.; Healy, J.; Saul, N.; and Grossberger, L. 2018. UMAP: Uniform Manifold Approximation and Projection. *The Journal of Open Source Software*, 3(29): 861.

Ming, Y.; Sun, Y.; Dia, O.; and Li, Y. 2023. How to Exploit Hyperspherical Embeddings for Out-of-Distribution Detection? In *ICLR*.

Mohseni, S.; Pitale, M.; Yadawa, J.; and Wang, Z. 2020. Self-Supervised Learning for Generalizable Out-of-Distribution Detection. In *AAAI*, volume 34, 5216–5223.

Morteza, P.; and Li, Y. 2022. Provable Guarantees for Understanding Out-of-Distribution Detection. In *AAAI*, volume 36, 7831–7840.

Netzer, Y.; Wang, T.; Coates, A.; Bissacco, A.; Wu, B.; Ng, A. Y.; et al. 2011. Reading Digits in Natural Images With Unsupervised Feature Learning. In *NeurIPS Workshop*, volume 2011, 7. Granada, Spain.

Rawat, W.; and Wang, Z. 2017. Deep Convolutional Neural Networks for Image Classification: A Comprehensive Review. *Neural Computation*, 29(9): 2352–2449.

Regmi, S.; Panthi, B.; Dotel, S.; Gyawali, P. K.; Stoyanov, D.; and Bhattarai, B. 2024a. T2FNORM: Train-Time Feature Normalization for OOD Detection in Image Classification. In *CVPR*, 153–162.

Regmi, S.; Panthi, B.; Ming, Y.; Gyawali, P. K.; Stoyanov, D.; and Bhattarai, B. 2024b. ReweightOOD: Loss Reweighting for Distance-Based OOD Detection. In *CVPR*, 131–141.

Ren, J.; Fort, S.; Liu, J.; Roy, A. G.; Padhy, S.; and Lakshminarayanan, B. 2021. A Simple Fix to Mahalanobis Distance for Improving Near-OOD Detection. *arXiv preprint arXiv:2106.09022*.

Ren, J.; Liu, P. J.; Fertig, E.; Snoek, J.; Poplin, R.; Depristo, M.; Dillon, J.; and Lakshminarayanan, B. 2019. Likelihood Ratios for Out-of-Distribution Detection. *NeurIPS*, 32.

Sehwag, V.; Chiang, M.; and Mittal, P. 2020. SSD: A Unified Framework for Self-supervised Outlier Detection. In *ICLR*.

Sun, Y.; Ming, Y.; Zhu, X.; and Li, Y. 2022. Out-of-Distribution Detection With Deep Nearest Neighbors. In *ICML*, 20827–20840. PMLR.

Szyc, K.; Walkowiak, T.; and Maciejewski, H. 2023. Why Out-of-Distribution Detection Experiments Are Not Reliable-Subtle Experimental Details Muddle the OOD Detector Rankings. In *UAI*, 2078–2088. PMLR.

Tack, J.; Mo, S.; Jeong, J.; and Shin, J. 2020. Csi: Novelty Detection Via Contrastive Learning on Distributionally Shifted Instances. *NeurIPS*, 33: 11839–11852.

Tao, L.; Du, X.; Zhu, J.; and Li, Y. 2023. Non-Parametric Outlier Synthesis. In *ICLR*.

Thulasidasan, S.; Chennupati, G.; Bilmes, J. A.; Bhattacharya, T.; and Michalak, S. 2019. On Mixup Training: Improved Calibration and Predictive Uncertainty for Deep Neural Networks. *NeurIPS*, 32.

Ulmer, D.; Meijerink, L.; and Cinà, G. 2020. Trust Issues: Uncertainty Estimation Does Not Enable Reliable OOD Detection on Medical Tabular Data. In *Machine Learning for Health*, 341–354. PMLR.

Van Horn, G.; Mac Aodha, O.; Song, Y.; Cui, Y.; Sun, C.; Shepard, A.; Adam, H.; Perona, P.; and Belongie, S. 2018. The Inaturalist Species Classification and Detection Dataset. In *CVPR*, 8769–8778.

Vishwakarma, H.; Lin, H.; and Vinayak, R. K. 2024. Taming False Positives in Out-of-Distribution Detection With Human Feedback. In *AISTATS*, 1486–1494. PMLR.

Wang, H.; Liu, W.; Bocchieri, A.; and Li, Y. 2021. Can Multi-Label Classification Networks Know What They Don't Know? *NeurIPS*, 34: 29074–29087.

Wang, Q.; Liu, F.; Zhang, Y.; Zhang, J.; Gong, C.; Liu, T.; and Han, B. 2022. Watermarking for Out-of-Distribution Detection. *NeurIPS*, 35: 15545–15557.

Wang, T.; and Isola, P. 2020. Understanding Contrastive Representation Learning Through Alignment and Uniformity on the Hypersphere. In *ICML*, 9929–9939. PMLR.

Wu, B.; Jiang, J.; Ren, H.; Du, Z.; Wang, W.; Li, Z.; Cai, D.; He, X.; Lin, B.; and Liu, W. 2023. Towards In-Distribution Compatible Out-of-Distribution Detection. In *AAAI*, volume 37, 10333–10341.

Xiao, J.; Hays, J.; Ehinger, K. A.; Oliva, A.; and Torralba, A. 2010. Sun Database: Large-Scale Scene Recognition From Abbey to Zoo. In *CVPR*, 3485–3492. IEEE.

Xu, P.; Ehinger, K. A.; Zhang, Y.; Finkelstein, A.; Kulkarni, S. R.; and Xiao, J. 2015. Turkergaze: Crowdsourcing Saliency With Webcam Based Eye Tracking. *arXiv preprint arXiv:1504.06755*.

Yang, J.; Wang, P.; Zou, D.; Zhou, Z.; Ding, K.; Peng, W.; Wang, H.; Chen, G.; Li, B.; Sun, Y.; et al. 2022. Openood: Benchmarking generalized out-of-distribution detection. *NeurIPS*, 35.

Yang, J.; Zhou, K.; Li, Y.; and Liu, Z. 2021. Generalized Out-of-Distribution Detection: A Survey. *arXiv preprint arXiv:2110.11334*.

Yu, F.; Seff, A.; Zhang, Y.; Song, S.; Funkhouser, T.; and Xiao, J. 2015. LSUN: Construction of a Large-Scale Image Dataset Using Deep Learning With Humans in the Loop. *arXiv preprint arXiv:1506.03365*.

Yun, S.; Han, D.; Oh, S. J.; Chun, S.; Choe, J.; and Yoo, Y. 2019. Cutmix: Regularization Strategy to Train Strong Classifiers With Localizable Features. In *ICCV*, 6023–6032.

Zhang, J.; Fu, Q.; Chen, X.; Du, L.; Li, Z.; Wang, G.; Han, S.; Zhang, D.; et al. 2022. Out-of-Distribution Detection Based on In-Distribution Data Patterns Memorization With Modern Hopfield Energy. In *ICLR*.

Zhou, B.; Lapedriza, A.; Khosla, A.; Oliva, A.; and Torralba, A. 2017. Places: A 10 Million Image Database for Scene Recognition. *IEEE T-PAMI*, 40(6): 1452–1464.