SIREN: Shaping Representations for Detecting Out-of-Distribution Objects

Xuefeng Du, Gabriel Gozum, Yifei Ming, Yixuan Li
Department of Computer Sciences
University of Wisconsin-Madison
{xfdu,gozum,alvinming,sharonli}@cs.wisc.edu

Abstract

Detecting out-of-distribution (OOD) objects is indispensable for safely deploying object detectors in the wild. Although distance-based OOD detection methods have demonstrated promise in image classification, they remain largely unexplored in object-level OOD detection. This paper bridges the gap by proposing a distance-based framework for detecting OOD objects, which relies on the model-agnostic representation space and provides strong generality across different neural architectures. Our proposed framework SIREN contributes two novel components: (1) a representation learning component that uses a trainable loss function to shape the representations into a mixture of von Mises-Fisher (vMF) distributions on the unit hypersphere, and (2) a test-time OOD detection score leveraging the learned vMF distributions in a parametric or non-parametric way. SIREN achieves competitive performance on both the recent detection transformers and CNN-based models, improving the AUROC by a large margin compared to the previous best method. Code is publicly available at https://github.com/deeplearning-wisc/siren.

1 Introduction

Teaching object detectors to be aware of out-of-distribution (OOD) data is indispensable for building reliable AI systems. Today, the mainstream object detection models have been operating in the closed-world setting. That is, a model will match an object to one of the given class labels, even if it is irrelevant. Instead, the open-world setting emphasizes that objects from the unknown classes can naturally emerge, which should not be blindly predicted into a known class. In safety-critical applications, such as autonomous driving, failing to detect OOD objects on the road can directly lead to disastrous accidents [50]. The situation can be better avoided if the object detector recognizes the object as unfamiliar and appropriately cautions the human driver to take over.

In this paper, we pioneer a distance-based framework for detecting OOD objects. Currently, the distance-based method remains largely unexplored in object-level OOD detection. In particular, by operating in the representation space, distance-based methods are model-agnostic and provide strong generality across neural architectures. In contrast, existing approaches derive highly specialized OOD detection scores based on the outputs of the object detectors, which may not be seamlessly applicable across architectures. For example, the classification output of the Faster R-CNN [52] is optimized by the multi-class softmax loss, whereas the recent transformer-based object detection networks such as DEFORMABLE-DETR [75] uses multi-label focal loss [36]. Thereby, while output-based OOD scoring functions may be limited to specific architectures, distance-based methods are not.

Although distance-based OOD scoring functions have been studied in image classification, they do not trivially transfer to object detection models. For example, [33] modeled the feature embedding space as a mixture of multivariate Gaussian distributions and used the maximum Mahalanobis distance [41] to all class centroids for OOD detection. However, we observe that the modern object detection

36th Conference on Neural Information Processing Systems (NeurIPS 2022).

(a) Vanilla DETR (b) SIREN (Ours)

Figure 1: (a) Feature embeddings from the penultimate layer of a vanilla DEFORMABLE-DETR [75] trained on the PASCAL-VOC dataset [14], which display irregular distributions. (b) Feature embeddings shaped by the proposed SIREN, which form compact clusters on the unit hypersphere.

models such as DEFORMABLE-DETR [75] produce highly irregular embeddings (Figure 1 (a)), which do not fit the Gaussian distributional assumption. As a result, the OOD detection score relying on such suboptimal embeddings can misbehave.

We propose a novel framework called SIREN, tackling two highly dependent problems—representation learning and OOD detection—in one synergistic framework. Concretely, SIREN contributes two novel components: **(1)** We introduce an end-to-end trainable loss that enables **S**hap**I**ng the **R**epres**EN**tations into a desired parametric form (Section 3.1). In particular, we model the representations by the von Mises-Fisher (vMF) distribution, a classic probability distribution in directional statistics for hyperspherical data with the unit norm. Our loss function encourages the normalized embedding to be aligned with its class prototype and shapes the overall representations into compact clusters for each class. Compared to the Gaussian distribution, using the vMF distribution avoids estimating large covariance matrices for high-dimensional data that is shown to be costly and unstable [7, 65]. **(2)** We explore test-time OOD detection by leveraging the optimized embeddings in a parametric or non-parametric way (Section 3.2). We propose a new test-time OOD score based on the learned class-conditional vMF distributions. The parameterization of the vMF distribution is directly obtainable after training, without requiring separate estimation. Different from Mahalanobis distance [33], the proposed parametric score in principle suits the learned vMF distributions on the hypersphere. Additionally, we explore a non-parametric nearest neighbor distance for OOD detection [60], which is agnostic to the type of distribution of the feature space.

Empirically, SIREN establishes superior performance on both transformer-based and CNN-based models. On PASCAL-VOC, SIREN outperforms the latest baseline OW-DETR [18] by a significant margin (↑22.53% in AUROC). Moreover, our framework is model-agnostic and does not incur changes to the existing network architecture. The proposed loss can be flexibly added as a plug-in module on top of modern architectures, as we show in Section 4.

Our key contributions are summarized as follows:

1. To the best of our knowledge, SIREN pioneers a distance-based approach for object-level OOD detection. Different from previous works, SIREN does not rely on specialized output-based OOD scores, and can generalize across different architectures in a model-agnostic fashion.

2. SIREN establishes competitive results on a challenging object-level OOD detection task. Compared to the latest method [18], SIREN improves the OOD detection performance by a considerable margin while preserving the mAP on the ID task. We show that SIREN is effective for both recent transformer-based and classic CNN-based models.

3. We shape representations via a novel vMF-based formulation for object-level OOD detection. We conduct in-depth ablations to understand how different factors impact the performance of SIREN (Section 5).

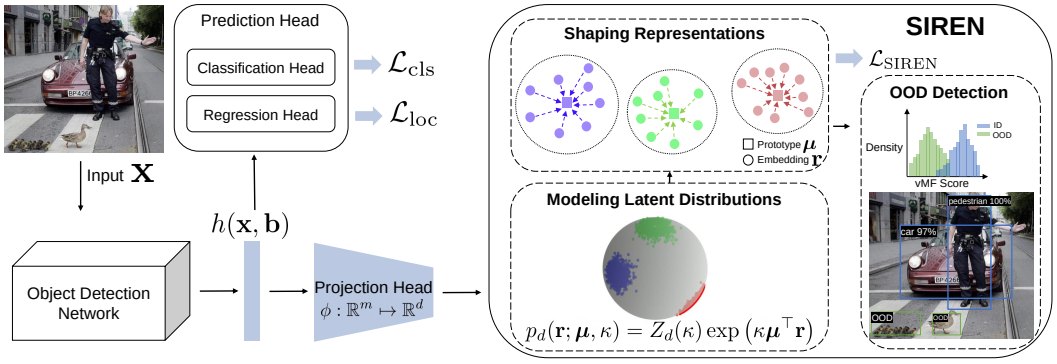

Figure 2: **Overview of the proposed learning framework SIREN.** We introduce a new loss $\mathcal{L}_{\text{SIREN}}$ which shapes the representations on the unit hypersphere into compact class-conditional vMF distributions. The embedding $\mathbf{r} \in \mathbb{R}^d$ has unit norm $\|\mathbf{r}\|^2 = 1$. In testing, we can employ either parametric or non-parametric distance functions for OOD detection. See Section 3 for details.

2 Preliminaries: Object-level OOD Detection

We start by introducing the OOD detection problem for object detection in the open-world setting, which has received increasing research attention lately [13, 18]. Our goal is to train object detection networks that can simultaneously: (1) localize and classify objects belonging to known categories accurately, and (2) identify unfamiliar objects outside the training categories. Compared to image-level OOD detection, object-level OOD detection is more suitable for real-world machine learning systems, yet also more challenging as it requires reasoning OOD uncertainty at the fine-grained object level. Since natural images are composed of multiple objects, knowing which regions of an image are anomalous allows for safe handling of unfamiliar objects.

Notations. We denote the input and label space by $\mathcal{X} = \mathbb{R}^q$ and $\mathcal{Y} = \{1, 2, ..., C\}$, respectively. Let $\mathbf{x} \in \mathcal{X}$ be the input image, $\mathbf{b} \in \mathbb{R}^4$ be the bounding box coordinates associated with objects in the image, and $y \in \mathcal{Y}$ be the semantic label of the object. An object detection model is trained on ID dataset $\mathcal{D}_{\text{tr}}^{\text{in}} = \{(\mathbf{x}_i, \mathbf{b}_i, y_i)\}_{i=1}^{M}$ drawn from an unknown joint distribution $\mathcal{P}$. We use neural networks with parameters θ to model the bounding box regression $p_\theta(\mathbf{b}|\mathbf{x})$ and the classification $p_\theta(y|\mathbf{x}, \mathbf{b})$.

Object-level OOD detection. The OOD detection can be formulated as a binary classification problem, distinguishing between the in- vs. out-of-distribution objects. Let $P_{\mathcal{X}}$ denote the marginal probability distribution on $\mathcal{X}$. Given a test input $\mathbf{x}' \sim P_{\mathcal{X}}$, as well as an object $\mathbf{b}'$ predicted by the object detector, the goal is to predict a binary outcome $g(\mathbf{x}', \mathbf{b}')$. We use $g = 1$ to indicate a detected object being ID, and $g = 0$ being OOD, with semantics outside the support of $\mathcal{Y}$.

3 Proposed Method

Overview. Our framework SIREN is illustrated in Figure 2, which trains an object detector in tandem with a representation-shaping branch. The object detector backbone $f : \mathcal{X} \mapsto \mathbb{R}^m$ maps an object to its feature embedding $h(\mathbf{x}, \mathbf{b}) \in \mathbb{R}^m$ (often referred to as the penultimate layer). In addition, we introduce a new MLP projection head $\phi : \mathbb{R}^m \mapsto \mathbb{R}^d$ that maps the $h(\mathbf{x}, \mathbf{b})$ to a lower-dimensional embedding $\mathbf{r} \in \mathbb{R}^d$ ($d < m$) with unit norm $\|\mathbf{r}\|^2 = 1$. The normalized embeddings are also referred to as *hyperspherical embeddings*, since they are on a unit hypersphere. In designing SIREN, we address two key challenges: **(1)** How to shape the hyperspherical representations into desirable probability distributions during training time (Section 3.1)? **(2)** How to perform test-time OOD detection by leveraging the learned distributions (Section 3.2)? Our method does not incur any change to the object detection network backbone. The proposed regularization can be flexibly used as a plug-in module on top of the modern architectures, as we will show in Section 4.

3.1 SIREN: Shaping Representations

Modeling the latent distributions. We propose to model the latent representations by the von Mises-Fisher (vMF) distribution [43], a probability distribution in directional statistics for spherical data with unit norm $\|\mathbf{r}\|^2 = 1$. The probability density function for a unit vector $\mathbf{r}$ in $\mathbb{R}^d$ is given as follows:

$$p_d(\mathbf{r}; \boldsymbol{\mu}, \kappa) = Z_d(\kappa) \exp\left(\kappa \boldsymbol{\mu}^\top \mathbf{r}\right), \tag{1}$$

where $\kappa \geq 0$, $\|\boldsymbol{\mu}\|^2 = 1$, and the normalization factor $Z_d(\kappa)$ is defined as:

$$Z_d(\kappa) = \frac{\kappa^{d/2-1}}{(2\pi)^{d/2} I_{d/2-1}(\kappa)}, \tag{2}$$

where I_v is the modified Bessel function of the first kind with order v. $Z_d(\kappa)$ can be calculated in closed form based on κ and the dimensionality d. Importantly, the vMF distribution is characterized by two parameters: the mean vector $\boldsymbol{\mu}$ and concentration parameter κ. Samples that are more aligned with the center $\boldsymbol{\mu}$ have a higher probability density, and vice versa. Here κ indicates the tightness of the distribution around the mean direction $\boldsymbol{\mu}$. The larger value of κ, the stronger the distribution is concentrated in the mean direction. In the extreme case of $\kappa = 0$, the sample points are distributed uniformly on the hypersphere.

When considering multiple classes, we can model the embedding space as a mixture of class-conditional vMF distributions, one for each class $c \in \{1, 2, ..., C\}$:

$$p_d^c(\mathbf{r}; \boldsymbol{\mu}_c, \kappa_c) = Z_d(\kappa_c) \exp\left(\kappa_c \boldsymbol{\mu}_c^\top \mathbf{r}\right), \tag{3}$$

where κ_c and $\boldsymbol{\mu}_c$ are class-conditional parameters. Under this probability model, an embedding vector $\mathbf{r}$ is assigned to class c with the following normalized probability:

$$p(y = c|\mathbf{r}; \{\kappa_j, \boldsymbol{\mu}_j\}_{j=1}^C) = \frac{Z_d\left(\kappa_c\right) \exp\left(\kappa_c \boldsymbol{\mu}_c^\top \mathbf{r}\right)}{\sum_{j=1}^C Z_d\left(\kappa_j\right) \exp\left(\kappa_j \boldsymbol{\mu}_j^\top \mathbf{r}\right)}. \tag{4}$$

Shaping representations. Our key idea is to design an end-to-end trainable loss function that enables **S**hap**I**ng the **R**epres**EN**tations into a mixture of vMF distributions, which facilitates test-time OOD detection in the representation space (Section 3.2). We therefore name our method SIREN. The learned mapping function projects an input to a point in the embedding space, where higher probability is assigned to the correct class in comparison to incorrect classes. To achieve this, we can perform maximum likelihood estimation (MLE) on the training data:

$$\text{argmax}_\theta \prod_{i=1}^{M} p(y_i|\mathbf{r}_i; \{\kappa_j, \boldsymbol{\mu}_j\}_{j=1}^C), \tag{5}$$

where i is the index of the object embedding and M is the size of the training set. By taking the negative log-likelihood, the objective function is equivalent to minimizing the following loss:

$$\mathcal{L}_{\text{SIREN}} = -\frac{1}{M} \sum_{i=1}^{M} \log \frac{Z_d\left(\kappa_{y_i}\right) \exp\left(\kappa_{y_i} \boldsymbol{\mu}_{y_i}^\top \mathbf{r}_i\right)}{\sum_{j=1}^C Z_d\left(\kappa_j\right) \exp\left(\kappa_j \boldsymbol{\mu}_j^\top \mathbf{r}_i\right)}, \tag{6}$$

where y_i is the ground truth label for the embedding $\mathbf{r}_i$. In effect, $\mathcal{L}_{\text{SIREN}}$ encourages the object embeddings to be aligned with its class prototype, which shapes the representations such that objects in each class form a compact cluster on the hypersphere; see Figure 1 (b).

Prototype estimation and update. During training, SIREN estimates the class-conditional object prototypes $\boldsymbol{\mu}_c, c \in \{1, 2, ..., C\}$. The conventional approach for estimating the prototypes is to calculate the mean vector of all training samples (or a subset of them) for each class, and update it periodically during training [74]. Despite its simplicity, this method requires alternating training and prototype estimation, which incurs a heavy computational toll and causes undesirable latency. Instead, we update the class-conditional prototypes in an exponential-moving-average (EMA) manner [34, 63]:

$$\boldsymbol{\mu}_c := \text{Normalize}(\alpha \boldsymbol{\mu}_c + (1 - \alpha)\mathbf{r}), \forall c \in \{1, 2, \ldots, C\}, \tag{7}$$

Algorithm 1 SIREN: Shaping Representations for object-level OOD detection

Input: ID training data $\mathcal{D}_{\text{tr}}^{\text{in}} = \{(\mathbf{x}_i, \mathbf{b}_i, y_i)\}_{i=1}^{M}$, randomly initialized object detector and MLP projection head with parameter θ, loss weight β for $\mathcal{L}_{\text{SIREN}}$, and learnable $\{\kappa_c\}_{c=1}^{C}$.
Output: Object detector with parameter θ^* and OOD detector G.
while *train* **do**
  1. Update class-conditional prototypes $\boldsymbol{\mu}_c$ with the hyperspherical embeddings $\mathbf{r}$ by Equation (7).
  2. Calculate the vMF-based representation shaping loss $\mathcal{L}_{\text{SIREN}}$ by Equation (6).
  3. Update the learnable $\{\kappa_c\}_{c=1}^{C}$ and the network parameters θ using Equation (8).

end
while *eval* **do**
  1. Calculate the OOD score by Equation (9) or Equation (11).
  2. Perform OOD detection by Equation (10).

end

where α is the prototype update factor, and $\mathbf{r}$ denotes the normalized object embeddings from class c. The update can be done efficiently with negligible cost, enabling end-to-end training.

Overall training objective. The overall training objective combines the standard object detection loss, along with our new representation shaping loss $\mathcal{L}_{\text{SIREN}}$:

$$\min_{\theta,\kappa} \mathbb{E}_{(\mathbf{x},\mathbf{b},y)\sim\mathcal{P}} \left[\mathcal{L}_{\text{cls}} + \mathcal{L}_{\text{loc}}\right] + \beta \cdot \mathcal{L}_{\text{SIREN}}, \tag{8}$$

where β is the weight of our representation shaping loss. $\mathcal{L}_{\text{cls}}$ and $\mathcal{L}_{\text{loc}}$ are losses for classification and bounding box regression, respectively. We provide extensive empirical evidence in Section 4 demonstrating the efficacy of our loss function.

To the best of our knowledge, our work makes the first attempt to explore vMF-based learning and inference for object-level OOD detection. To highlight the novelty of the loss itself: we introduce a novel *learnable* $\{\kappa_j\}_{j=1}^{C}$ for each ID class in Equation (6), instead of using fixed values. Our loss allows concentration parameter κ to adaptively and flexibly capture the class-conditional feature statistics during training. This is desirable when each ID class may have its own concentration in the hypersphere. We will later show that such learnable $\{\kappa_j\}_{j=1}^{C}$ enables better OOD detection performance (Section 5).

3.2 Test-time OOD Detection

During inference, we explore and contrast two types of uncertainty scores for detecting OOD objects.

Parametric vMF score. We propose a new test-time OOD score based on the learned class-conditional vMF distributions, parameterized by $\{\widehat{\kappa}_c, \boldsymbol{\mu}_c\}_{c=1}^{C}$. Here $\widehat{\kappa}_c$ denotes the learned concentration parameter for class c, which captures the concentration of representations for class c. For a test-time object $(\mathbf{x}', \mathbf{b}')$, we use the largest estimated class-conditional likelihood as the OOD score:

$$S\left(\mathbf{x}', \mathbf{b}'\right) = \max_{c} Z_d(\widehat{\kappa}_c) \exp\left(\widehat{\kappa}_c \boldsymbol{\mu}_c^\top \mathbf{r}'\right), \tag{9}$$

where $\mathbf{r}' = \phi(h(\mathbf{x}', \mathbf{b}'))$ is the normalized embedding from the MLP projection head. Our OOD detection score thus in principle suits our learned embeddings and vMF distributions. For OOD detection, one can use the level set to distinguish between ID and OOD objects:

$$G\left(\mathbf{x}', \mathbf{b}'\right) = \begin{cases} 1 & \text{if } S\left(\mathbf{x}', \mathbf{b}'\right) \geq \gamma \\ 0 & \text{if } S\left(\mathbf{x}', \mathbf{b}'\right) < \gamma \end{cases} \tag{10}$$

The threshold γ can be chosen so that a high fraction of ID data (e.g., 95%) is correctly classified. For objects classified as ID, one can obtain the bounding box and class prediction using the prediction head as usual.

Non-parametric KNN score. To relax the distributional assumption on the learned embeddings, we additionally employ a non-parametric KNN distance for OOD detection, which performs well on compact and normalized feature space. Following Sun *et al.* [60], the KNN distance is defined as:

$$S\left(\mathbf{x}', \mathbf{b}', k\right) = -\left\|\mathbf{r}' - \mathbf{r}_{(k)}\right\|_2, \tag{11}$$

where $\mathbf{r}_{(k)}$ denotes the normalized embedding of the k-th nearest neighbor (in the training data), for the test embedding $\mathbf{r}'$. Our algorithm is summarized in Algorithm 1.

Remark 1. *Different from Mahalanobis distance [33], our parametric vMF-based OOD detection score operates under the same distributional model as the training process, and hence enjoys mathematical compatibility. Computationally, we can directly use the parameters of vMF distributions (such as κ) learned from training. In contrast, the Mahalanobis distance requires a separate test-time estimation of feature statistics—which involves an expensive and numerically unstable step of calculating the covariance matrix. The distinction of the uncertainty surface calculated by both our vMF score and the Mahalanobis distance is qualitatively demonstrated in Figure 3. The data points are sampled from a mixture of three class-conditional vMF distributions (see details in Appendix F).*

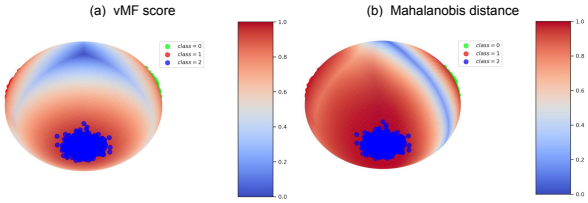

Figure 3: The uncertainty surface is calculated using our vMF score (a) and the Mahalanobis distance (b). We showcase one class for visual clarity.

4 Experiments

In this section, we validate the effectiveness of SIREN on object detection models, including the latest transformer-based (Section 4.1) and flagship CNN-based models (Section 4.2).

Datasets. Following [13], we use PASCAL-VOC[1] [14] and Berkeley DeepDrive (BDD100K)[2] [73] datasets as the ID training data. For both tasks, we evaluate on two OOD datasets that contain a subset of images from: MS-COCO [37] and OPENIMAGES (validation set) [30]. Extensive details on the datasets are in Appendix A.

Metrics. For evaluating the OOD detection performance, we report: (1) the false positive rate (FPR95) of OOD objects when the true positive rate of ID samples is at 95%; (2) the area under the receiver operating characteristic curve (AUROC). For evaluating the object detection performance on the ID task, we report the common metric mAP.

4.1 Evaluation on Transformer-based Model

Experimental details. We adopt a very recent DEFORMABLE-DETR (DDETR) architecture [75]. DDETR introduces multi-scale deformable attention modules in the transformer encoder and decoder layers of DETR [4], and provides better convergence and lower complexity. The multi-scale feature maps in DDETR are extracted from a ResNet-50 [21] pre-trained on ImageNet in a self-supervised fashion, *i.e.*, DINO [5]. We use the embeddings at the penultimate layer of the decoder in DDETR for projection. For the projection head, we use a two-layer MLP with a ReLU nonlinearity, with dimensionality $256 \rightarrow d \rightarrow d$. The dimension d of the unit hypersphere is 16 for PASCAL-VOC and 64 for BDD100K. The default weight β for the SIREN is 1.5 and the prototype update factor α is 0.95. We initialize the learnable κ to be 10 for all classes. The k in the KNN distance is set to 10. Ablations on the hyperparameters are provided in Section 5 and Appendix C. Other hyperparameters are the same as the default ones in DDETR [75].

SIREN achieves superior performance. In Table 1, we compare SIREN with competitive OOD detection methods in literature. For a fair comparison, all the methods only use ID data for training. SIREN outperforms competitive baselines, including Mahalanobis distance [33], KNN distance [60], CSI [61], Gram matrices [53] and Dismax [39]. These baselines operate on feature embedding space, allowing a fair comparison. Note that other common output-based methods (such as MSP [23], ODIN [35], and energy [38]) are not directly applicable for multi-label classification networks in DDETR. For these methods relying on a multi-class classification model, we will later provide comparisons on the Faster R-CNN model in Section 4.2. Implementation details and the training time for all the baseline methods are reported in Appendix D and E. We highlight a few observations:

1) The comparison between **SIREN vs. Mahalanobis** highlights precisely the benefits of our embedding shaping loss $\mathcal{L}_{\text{SIREN}}$. As shown in Figure 1, the vanilla DETR model produces ill-

[1]PASCAL-VOC consists of the following ID labels: Person, Car, Bicycle, Boat, Bus, Motorbike, Train, Airplane, Chair, Bottle, Dining Table, Potted Plant, TV, Sofa, Bird, Cat, Cow, Dog, Horse, Sheep.

[2]BDD100K consists of ID labels: Pedestrian, Rider, Car, Truck, Bus, Train, Motorcycle, Bicycle, Traffic light, Traffic sign.

In-distribution dataset	Method	FPR95 ↓	AUROC ↑	mAP (ID)↑
		OOD: MS-COCO / OpenImages		
PASCAL-VOC	Mahalanobis [33]	97.39 / 97.88	50.28 / 49.08	60.6
	Gram matrices [53]	94.16 / 95.29	43.97 / 38.81	60.6
	KNN [60]	91.80 / 91.36	62.15 / 59.64	60.6
	CSI [61]	84.00 / 79.16	55.07 / 51.37	59.5
	VOS [13]	97.46 / 97.07	54.40 / 52.77	60.3
	OW-DETR [18]	93.09 / 93.82	55.70 / 57.80	58.3
	Dismax [39]	82.05 / 76.37	75.21 / 70.66	60.1
	SIREN-vMF (ours)	75.49±0.8 / 78.36±1.0	76.10±0.1 / 71.05±0.1	60.8±0.1
	SIREN-KNN (ours)	**64.77**±0.2 / **65.99**±0.5	**78.23**±0.2 / **74.93**±0.1	60.8±0.1
BDD100K	Mahalanobis [33]	70.86 / 71.43	76.83 / 77.98	31.3
	Gram matrices [53]	73.81 / 71.56	60.13 / 57.14	31.3
	KNN [60]	64.75 / 61.13	80.90 / 79.64	31.3
	CSI [61]	70.27 / 71.30	77.93 / 76.42	29.9
	VOS [13]	76.44 / 72.58	77.33 / 76.62	31.0
	OW-DETR [18]	80.78 / 77.37	70.29 / 73.78	28.1
	Dismax [39]	77.62 / 81.23	72.14 / 67.18	31.2
	SIREN-vMF (ours)	67.54±1.3 / 66.31±0.9	80.06±0.5 / 79.77±1.2	31.3±0.0
	SIREN-KNN (ours)	**53.97**±0.7 / **47.28**±0.3	**86.56**±0.1 / **89.00**±0.4	31.3±0.0

Table 1: **Main results.** Comparison with competitive out-of-distribution detection methods. All baseline methods are based on the same model backbone DDETR. ↑ indicates larger values are better and ↓ indicates smaller values are better. All values are percentages. **Bold** numbers are superior results. We report standard deviations estimated across 3 runs. SIREN-vMF/KNN denotes using vMF score and KNN distance during inference.

conditioned embeddings that do not conform to multivariate Gaussian distributions, rendering the Mahalanobis approach ineffective (with AUROC around 50%–which is random guessing). In contrast, SIREN-vMF improves the OOD detection performance (AUROC) by **25.82**% on PASCAL-VOC (MS-COCO as OOD). Different from the Mahalanobis distance, our parametric vMF scoring function naturally suits the learned hyperspherical embeddings with vMF distributions. The advantage of the vMF loss can be further verified by observing that SIREN-KNN outperforms directly applying KNN distance on the vanilla DDETR (**16.08**% AUROC improvement on VOC with COCO as OOD).

2) SIREN outperforms the latest methods VOS [13] and OW-DETR [18], which are designed for object detection models and serve as strong baselines for us. Compared with OW-DETR, SIREN-KNN substantially improves the AUROC by **22.53**% on PASCAL-VOC (COCO as OOD). OW-DETR uses the unmatched object queries with high confidence as the unknowns, and trains a binary classifier to separate ID and unknown objects. However, the unmatched object queries might be distributionally too close to the ID classes and thus displays limited improvement for OOD detection. In addition, VOS synthesizes virtual outliers from the class-conditional Gaussian distributions of the penultimate layer but fails to perform well due to the ill-conditioned embedding distribution in DDETR (non-Gaussian).

4.2 Evaluation on CNN-based Model

Going beyond detection transformers, we show that SIREN is also suitable and effective on CNN-based object-level OOD detection models, *e.g.*, Faster R-CNN [52]. Table 2 showcases the OOD detection performance with SIREN trained on PASCAL-VOC dataset and evaluated on both MS-COCO and OPENIMAGES datasets. In addition to baselines considered in Table 1, we include common output-based methods relying on multi-class classification, such as MSP, ODIN, and energy score.

In Table 2, we additionally report training time comparison. SIREN consistently improves OOD detection performance on both OOD datasets. Notably, SIREN performs better than

Method	FPR95 ↓	AUROC ↑	Time
	COCO/ OpenImages		
MSP	70.99 / 73.13	83.45 / 81.91	2.1 h
ODIN	59.82 / 63.14	82.20 / 82.59	2.1 h
Mahalanobis	96.46 / 96.27	59.25 / 57.42	2.1 h
Energy score	56.89 / 58.69	83.69 / 82.98	2.1 h
Gram matrices	62.75 / 67.42	79.88 / 77.62	2.1 h
KNN	52.67 / 53.67	87.14 / 84.54	2.1 h
CSI	59.91 / 57.41	81.83 / 82.95	4.9 h
GAN-synthesis	60.93 / 59.97	83.67 / 82.67	3.7 h
VOS	47.53 / 51.33	88.70 / 85.23	4.3 h
Dismax	84.38 / 86.93	74.56 / 71.53	2.2 h
SIREN-vMF (ours)	64.68 / 68.53	85.36 / 82.78	2.1 h
SIREN-KNN (ours)	**47.45 / 50.38**	**89.67 / 88.80**	2.1 h

Table 2: OOD detection results of SIREN and comparison with competitive baselines on two OOD datasets: COCO and OpenImages.

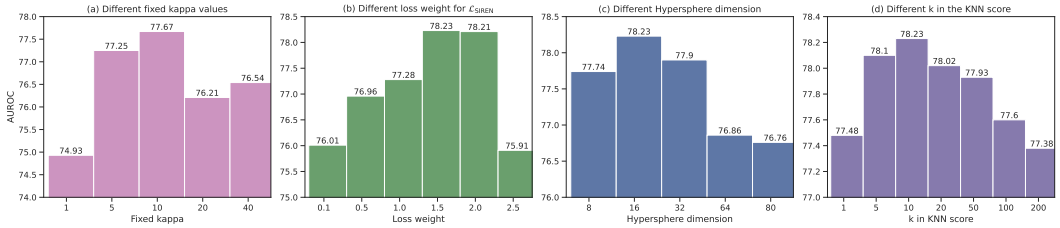

Figure 4: (a) Ablation study on using different fixed values of concentration parameter κ in Equation (12). (b) Ablation on different weights β for $\mathcal{L}_{\text{SIREN}}$ in Equation (8). (c) Ablation on the dimension of the hyperspherical embeddings $\mathbf{r}$. (d) Ablation study on the parameter k in the KNN-based OOD detection score for SIREN. Numbers are AUROC. The ID training dataset is PASCAL-VOC and OOD dataset is MS-COCO.

the previous best OOD detection approach VOS on Faster R-CNN while preserving the same training time as vanilla Faster R-CNN.

5 Ablations and Discussions

In this section, we provide ablation results on how different factors impact the performance of SIREN. For consistency, we present the analyses below based on the DDETR model. Unless otherwise pointed, we use the KNN distance by default.

Ablations on learnable vs. fixed κ. In this ablation, we show that our approach using learnable concentration parameters $\{\kappa_c\}_{c=1}^{C}$ is better than using the fixed ones. Our method with learnable κ is desirable, since each ID class may have its own concentration in the hypersphere. With fixed κ, the loss function can be simplified as follows:

$$\mathcal{L} = -\frac{1}{M}\sum_{i=1}^{M}\log\frac{\exp\left(\kappa\boldsymbol{\mu}_{y_i}^{\top}\mathbf{r}_i\right)}{\sum_{j=1}^{C}\exp\left(\kappa\boldsymbol{\mu}_{j}^{\top}\mathbf{r}_i\right)}. \tag{12}$$

Empirically, we indeed observe that employing learnable $\{\kappa_c\}_{c=1}^{C}$ achieves better OOD detection performance, with AUROC 78.23% on PASCAL-VOC model (MS-COCO as OOD). In Figure 4 (a), we show SIREN's performance when trained under different fixed κ values. The best model under the fixed $\kappa = 10$ achieves an AUROC of 77.67%. Too small of κ value (*e.g.*, $\kappa = 1$) leads to almost uniform distributions and is therefore not desirable.

Ablations on the SIREN loss weight β. Figure 4 (b) reports the OOD detection results as we vary the weight β for the representation shaping loss $\mathcal{L}_{\text{SIREN}}$. The model is evaluated on the MS-COCO dataset as OOD. Overall a mild weight works well. Across all the β values considered, SIREN consistently outperforms the baseline OOD detection methods in Table 1. One can use our SIREN loss as an easy plug-in module, with minimal hyperparameter tuning.

Ablations on the dimension d of hypersphere. SIREN projects the object feature embeddings into a lower-dimensional hypersphere in $\mathbb{R}^d$, which allows tractable vMF estimation. Figure 4 (c) shows the effect the embedding dimension d on the OOD detection performance. We find that a lower dimension between 16 and 64 achieves favorable and stable performance. In the extreme case with dimension $d = 8$, the model suffers from considerable information loss and degraded performance. On the other hand, too large of d causes training instability, which is not desirable either.

Ablations on the uncertainty score. We perform ablation on two variants of the vMF-based OOD detection score (*c.f.* Equation (9)): using the learned $\{\kappa_c\}_{c=1}^{C}$ vs. approximately estimate the κ parameters directly from the converged embeddings. In literature, there are several established methods for approximating κ [43, 55]. Denote $\bar{\mathbf{r}}_c$ the average of object embeddings for class c, the simplest approximate solution is given as follows:

$$\widehat{\kappa_c} = \frac{\|\bar{\mathbf{r}}_c\|(d - \|\bar{\mathbf{r}}_c\|^2)}{1 - \|\bar{\mathbf{r}}_c\|^2}, \tag{13}$$

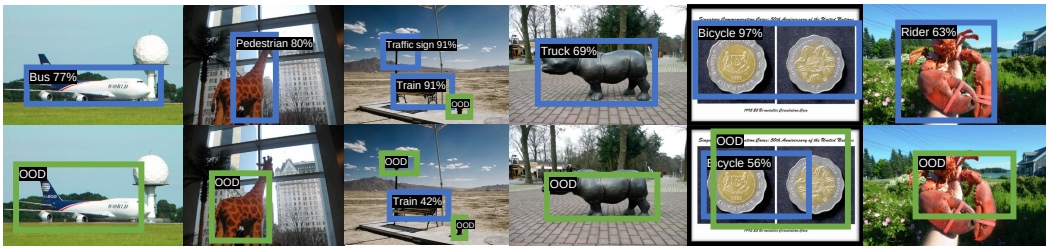

Figure 5: Visualization of detected objects on the OOD images (from MS-COCO and OPENIMAGES) by vanilla DDETR (*top*) and SIREN (*bottom*). The ID is BDD100K dataset. **Blue**: OOD objects classified as the ID classes. **Green**: OOD objects detected by DDETR and SIREN, which reduce false positives.

The proof is given in Appendix B. We show in Table 3 that using learned $\{\kappa_c\}_{c=1}^{C}$ avoids the imprecision in approximating $\widehat{\kappa_c}$ directly from embeddings. This affirms the importance of employing learnable concentration parameters in our SIREN loss. Moreover, the non-parametric KNN density estimation provides stronger flexibility and generality, and leads to better performance.

	FPR95 ↓	AUROC ↑
	COCO / OpenImages as OOD	
vMF w/ κ from [55]	78.60 / 78.42	73.03 / 70.27
vMF w/ learned κ (ours)	75.49 / 78.36	76.10 / 71.05
Non-parametric KNN distance	**64.77 / 65.99**	**78.23 / 74.93**

Table 3: Ablation on different OOD detection scores. The ID dataset is PASCAL-VOC.

Ablations on the projection head. We ablate on the nonlinearity in the projection head by comparing with SIREN trained with a linear layer in Table 4. The result shows using a nonlinear mapping for projection helps obtain a more expressive hypersphere, which improves OOD detection by 4.37% in terms of AUROC (MS-COCO as OOD, vMF as the OOD score).

	FPR95 ↓	AUROC ↑
	COCO / OpenImages as OOD	
vMF w/o nonlinearity	82.85 / 83.69	71.73 / 66.82
vMF w/ nonlinearity	**75.49 / 78.36**	**76.10 / 71.05**
KNN w/o nonlinearity	67.03 / 72.11	78.02 / 72.42
KNN w/ nonlinearity	**64.77 / 65.99**	**78.23 / 74.93**

Table 4: Ablation on the projection head. The ID dataset is PASCAL-VOC.

6 Qualitative analysis

In Figure 5, we visualize the predictions on several OOD images, using object detection models trained without SIREN (top) and with SIREN (bottom), respectively. The in-distribution data is BDD100K. SIREN better identifies OOD objects (in green) compared to a vanilla object detector DDETR, reducing false positives. Moreover, the confidence score of the false-positive objects of SIREN is lower than that of the vanilla model (see the train/bicycle in the 3rd/5th column).

7 Related work

OOD detection for classification can be broadly categorized into post hoc and regularization-based approaches [72]. In [2], the OpenMax score is proposed for OOD detection based on the extreme value theory (EVT). Subsequent work [23] proposed a simple baseline using maximum softmax confidence

(MSP). Several improvements have been proposed, such as deep ensemble [31], ODIN [35], distance-based score [33, 51, 53, 60], energy-based score [38, 49, 64], DICE [59] and ReAct score [58]. While post hoc methods often considered OOD scoring function alone, our framework considers both representation learning (at training) and OOD detection (at testing).

Along the line of regularization-based approaches, most prior works focused on regularizing the model's *output* [3, 16, 22, 26, 28, 42, 44, 47, 48, 62, 66, 71]. For example, the model is regularized to produce lower confidence [24, 32] or higher energy [13, 38] on the outlier data. In contrast, SIREN regularizes the model by shaping the *representations* during training. [39, 40, 61] employed self-supervised learning and learnable prototypes to learn desirable representations for novelty detection, respectively, but they used complex output-based score ensembles during inference, which are not generalizable and perform sub-optimally (see Tables 1 and 2). [57] shaped the latent representations as multivariate Gaussians but they trained a variational autoencoder and utilized the reconstruction error for open-set recognition. In contrast, SIREN directly shapes the embeddings of a *discriminative-based model*, and does not require training a generative model.

OOD detection for object detection is a rising topic with very few existing works. For Faster R-CNN, Du *et al.* [13] proposed to synthesize virtual outliers in the feature space for model regularization, which we compare in Section 4. [12] explored unknown-aware object detection by leveraging videos in the wild, whereas we focus on settings with still images only. For DETR, [18] adopted unmatched object queries that are with high confidence as unknowns, which did not focus on regularizing the model for desirable representations. Several works [9, 11, 19, 45, 46] used approximate Bayesian methods, such as MC-Dropout [15] for OOD detection. They require multiple inference passes to generate the uncertainty score, which are computationally expensive on larger datasets and models.

vMF distribution in ML has been adopted for supervised classification [29, 54], face verification [20], generative modeling [8], segmentation [25] and clustering [17], etc. Some works exploited vMF distribution for anomaly detection, they employed generative models and used the vMF distribution as the prior for zero-shot learning [6] and document analysis [76]. Both the problem settings and approaches are orthogonal to SIREN. To the best of our knowledge, our work makes the first attempt to employ vMF-based learning and inference with learnable parameters for object-level OOD detection. Metric learning has been explored for both classification [10, 67, 70, 74] and object detection [27, 56, 68, 69]. Unlike ours, they did not focus on the OOD detection problem.

8 Conclusion

In this paper, we propose a novel framework SIREN, which tackles object-level OOD detection with a distance-based approach. SIREN mitigates the key shortcoming of the previous output-based OOD detection approach, and explores a new vMF loss to shape representations for OOD detection. To the best of our knowledge, SIREN makes the first attempt to employ vMF-based learning and inference for OOD detection. SIREN establishes competitive performance on challenging object-level OOD detection tasks, evaluated broadly under both the recent detection transformers and CNN-based models. Our in-depth ablations provide further insights on the efficacy of SIREN. We hope our work inspires future research on OOD detection with representation shaping.

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 and releasing our code, we hope to raise stronger research and societal awareness towards the problem of out-of-distribution detection in real-world settings.

Acknowledgement

We thank Yiyou Sun and Ziyang (Jack) Cai for their valuable suggestions. The authors would also like to thank NeurIPS anonymous reviewers for their helpful feedback. The work is supported by the AFOSR Young Investigator Program Award.

References

[1] Arindam Banerjee, Inderjit S. Dhillon, Joydeep Ghosh, and Suvrit Sra. Clustering on the unit hypersphere using von mises-fisher distributions. *J. Mach. Learn. Res.*, 6:1345–1382, 2005.

[2] Abhijit Bendale and Terrance E Boult. Towards open set deep networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1563–1572, 2016.

[3] Petra Bevandić, Ivan Krešo, Marin Oršić, and Siniša Šegvić. Discriminative out-of-distribution detection for semantic segmentation. *arXiv preprint arXiv:1808.07703*, 2018.

[4] Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. End-to-end object detection with transformers. In *Proceedings of European Conference on Computer Vision*, pages 213–229, 2020.

[5] Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers. In *Proceedings of IEEE/CVF International Conference on Computer Vision*, pages 9630–9640, 2021.

[6] Xingyu Chen, Xuguang Lan, Fuchun Sun, and Nanning Zheng. A boundary based out-of-distribution classifier for generalized zero-shot learning. In *Proceedings of European Conference on Computer Vision*, pages 572–588, 2020.

[7] Xixian Chen, Michael R. Lyu, and Irwin King. Toward efficient and accurate covariance matrix estimation on compressed data. In *Proceedings of the International Conference on Machine Learning*, pages 767–776, 2017.

[8] Tim R. Davidson, Luca Falorsi, Nicola De Cao, Thomas Kipf, and Jakub M. Tomczak. Hyperspherical variational auto-encoders. In *Proceedings of the Thirty-Fourth Conference on Uncertainty in Artificial Intelligence*, pages 856–865, 2018.

[9] Kumari Deepshikha, Sai Harsha Yelleni, P. K. Srijith, and C. Krishna Mohan. Monte carlo dropblock for modelling uncertainty in object detection. *CoRR*, abs/2108.03614, 2021.

[10] Jiankang Deng, Jia Guo, Niannan Xue, and Stefanos Zafeiriou. Arcface: Additive angular margin loss for deep face recognition. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, pages 4690–4699, 2019.

[11] Akshay Raj Dhamija, Manuel Günther, Jonathan Ventura, and Terrance E. Boult. The overlooked elephant of object detection: Open set. In *Proceedings of IEEE Winter Conference on Applications of Computer Vision*, pages 1010–1019, 2020.

[12] Xuefeng Du, Xin Wang, Gabriel Gozum, and Yixuan Li. Unknown-aware object detection: Learning what you don't know from videos in the wild. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13678–13688, 2022.

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

[14] Mark Everingham, Luc Van Gool, Christopher K. I. Williams, John M. Winn, and Andrew Zisserman. The pascal visual object classes (VOC) challenge. *International Journal of Computer Vision*, 88(2):303–338, 2010.

[15] Yarin Gal and Zoubin Ghahramani. Dropout as a bayesian approximation: Representing model uncertainty in deep learning. In *Proceedings of International Conference on Machine Learning*, pages 1050–1059, 2016.

[16] Yonatan Geifman and Ran El-Yaniv. Selectivenet: A deep neural network with an integrated reject option. In *Proceedings of International Conference on Machine Learning*, pages 2151–2159. PMLR, 2019.

[17] Siddharth Gopal and Yiming Yang. Von mises-fisher clustering models. In *Proceedings of International Conference on Machine Learning*, pages 154–162, 2014.

[18] Akshita Gupta, Sanath Narayan, KJ Joseph, Salman Khan, Fahad Shahbaz Khan, and Mubarak Shah. Ow-detr: Open-world detection transformer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9235–9244, 2022.

[19] David Hall, Feras Dayoub, John Skinner, Haoyang Zhang, Dimity Miller, Peter Corke, Gustavo Carneiro, Anelia Angelova, and Niko Sünderhauf. Probabilistic object detection: Definition and evaluation. In *Proceedings of IEEE Winter Conference on Applications of Computer Vision*, pages 1020–1029, 2020.

[20] Md. Abul Hasnat, Julien Bohné, Jonathan Milgram, Stéphane Gentric, and Liming Chen. von mises-fisher mixture model-based deep learning: Application to face verification. *CoRR*, abs/1706.04264, 2017.

[21] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

[22] Matthias Hein, Maksym Andriushchenko, and Julian Bitterwolf. Why relu networks yield high-confidence predictions far away from the training data and how to mitigate the problem. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 41–50, 2019.

[23] Dan Hendrycks and Kevin Gimpel. A baseline for detecting misclassified and out-of-distribution examples in neural networks. *Proceedings of International Conference on Learning Representations*, 2017.

[24] Dan Hendrycks, Mantas Mazeika, and Thomas Dietterich. Deep anomaly detection with outlier exposure. In *Proceedings of International Conference on Learning Representations*, 2019.

[25] Jyh-Jing Hwang, Stella X. Yu, Jianbo Shi, Maxwell D. Collins, Tien-Ju Yang, Xiao Zhang, and Liang-Chieh Chen. Segsort: Segmentation by discriminative sorting of segments. In *Proceedings of IEEE/CVF International Conference on Computer Vision*, pages 7333–7343, 2019.

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

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

[28] Julian Katz-Samuels, Julia Nakhleh, Robert Nowak, and Yixuan Li. Training ood detectors in their natural habitats. In *Proceedings of International Conference on Machine Learning*, 2022.

[29] Takumi Kobayashi. T-vmf similarity for regularizing intra-class feature distribution. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6616–6625, 2021.

[30] Alina Kuznetsova, Hassan Rom, Neil Alldrin, Jasper Uijlings, Ivan Krasin, Jordi Pont-Tuset, Shahab Kamali, Stefan Popov, Matteo Malloci, Alexander Kolesnikov, et al. The open images dataset v4. *International Journal of Computer Vision*, pages 1–26, 2020.

[31] Balaji Lakshminarayanan, Alexander Pritzel, and Charles Blundell. Simple and scalable predictive uncertainty estimation using deep ensembles. In *Proceedings of Advances in Neural Information Processing Systems*, pages 6402–6413, 2017.

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

[33] Kimin Lee, Kibok Lee, Honglak Lee, and Jinwoo Shin. A simple unified framework for detecting out-of-distribution samples and adversarial attacks. In *Proceedings of Advances in Neural Information Processing Systems*, pages 7167–7177, 2018.

[34] Junnan Li, Caiming Xiong, and Steven C. H. Hoi. Mopro: Webly supervised learning with momentum prototypes. In *Proceedingd of International Conference on Learning Representations*, 2021.

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

[36] Tsung-Yi Lin, Priya Goyal, Ross B. Girshick, Kaiming He, and Piotr Dollár. Focal loss for dense object detection. *IEEE Trans. Pattern Anal. Mach. Intell.*, 42(2):318–327, 2020.

[37] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *Proceedings of European Conference on Computer Vision*, pages 740–755, 2014.

[38] Weitang Liu, Xiaoyun Wang, John Owens, and Yixuan Li. Energy-based out-of-distribution detection. *Proceedings of Advances in Neural Information Processing Systems*, pages 21464–21475, 2020.

[39] David Macêdo, Cleber Zanchettin, and Teresa Bernarda Ludermir. Distinction maximization loss: Efficiently improving out-of-distribution detection and uncertainty estimation simply replacing the loss and calibrating. *CoRR*, abs/2205.05874, 2022.

[40] David Macêdo, Tsang Ing Ren, Cleber Zanchettin, Adriano L. I. Oliveira, and Teresa Ludermir. Entropic out-of-distribution detection: Seamless detection of unknown examples. *IEEE Transactions on Neural Networks and Learning Systems*, 33(6):2350–2364, 2022.

[41] Prasanta Chandra Mahalanobis. On the generalized distance in statistics. National Institute of Science of India, 1936.

[42] Andrey Malinin and Mark Gales. Predictive uncertainty estimation via prior networks. In *Proceedings of Advances in Neural Information Processing Systems*, pages 7047–7058, 2018.

[43] Kanti V Mardia, Peter E Jupp, and KV Mardia. *Directional statistics*, volume 2. Wiley Online Library, 2000.

[44] Alexander Meinke and Matthias Hein. Towards neural networks that provably know when they don't know. In *Proceedings of International Conference on Learning Representations*, 2020.

[45] Dimity Miller, Feras Dayoub, Michael Milford, and Niko Sünderhauf. Evaluating merging strategies for sampling-based uncertainty techniques in object detection. In *Proceedings of International Conference on Robotics and Automation*, pages 2348–2354, 2019.

[46] Dimity Miller, Lachlan Nicholson, Feras Dayoub, and Niko Sünderhauf. Dropout sampling for robust object detection in open-set conditions. In *Proceedings of IEEE International Conference on Robotics and Automation*, pages 1–7, 2018.

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

[48] Sina Mohseni, Mandar Pitale, J. B. S. Yadawa, and Zhangyang Wang. Self-supervised learning for generalizable out-of-distribution detection. In *Proceedings of AAAI Conference on Artificial Intelligence*, pages 5216–5223, 2020.

[49] Peyman Morteza and Yixuan Li. Provable guarantees for understanding out-of-distribution detection. *Proceedings of the AAAI Conference on Artificial Intelligence*, 2022.

[50] Julia Nitsch, Masha Itkina, Ransalu Senanayake, Juan I. Nieto, Max Schmidt, Roland Siegwart, Mykel J. Kochenderfer, and Cesar Cadena. Out-of-distribution detection for automotive perception. In *Proceedings of IEEE International Intelligent Transportation Systems Conference*, pages 2938–2943, 2021.

[51] Jie Ren, Stanislav Fort, Jeremiah Liu, Abhijit Guha Roy, Shreyas Padhy, and Balaji Lakshminarayanan. A simple fix to mahalanobis distance for improving near-ood detection. *CoRR*, abs/2106.09022, 2021.

[52] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. In *Proceedings of Advances in Neural Information Processing Systems*, pages 91–99, 2015.

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

[54] Tyler R. Scott, Andrew C. Gallagher, and Michael C. Mozer. von mises-fisher loss: An exploration of embedding geometries for supervised learning. In *Proceedings of IEEE/CVF International Conference on Computer Vision*, pages 10592–10602, 2021.

[55] Suvrit Sra. A short note on parameter approximation for von mises-fisher distributions: and a fast implementation of $I_{\ s}(x)$. *Comput. Stat.*, 27(1):177–190, 2012.

[56] Bo Sun, Banghuai Li, Shengcai Cai, Ye Yuan, and Chi Zhang. FSCE: few-shot object detection via contrastive proposal encoding. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, pages 7352–7362, 2021.

[57] Xin Sun, Zhenning Yang, Chi Zhang, Keck Voon Ling, and Guohao Peng. Conditional gaussian distribution learning for open set recognition. In *Proceedings of IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13477–13486, 2020.

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

[59] Yiyou Sun and Yixuan Li. Dice: Leveraging sparsification for out-of-distribution detection. In *Proceedings of European Conference on Computer Vision*, 2022.

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

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

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

[63] Haobo Wang, Ruixuan Xiao, Yixuan Li, Lei Feng, Gang Niu, Gang Chen, and Junbo Zhao. Pico: Contrastive label disambiguation for partial label learning. In *Proceedings of International Conference on Learning Representations*, 2022.

[64] Haoran Wang, Weitang Liu, Alex Bocchieri, and Yixuan Li. Can multi-label classification networks know what they don't know? In *Advances in Neural Information Processing Systems*, 2021.

[65] Tongzhou Wang and Phillip Isola. Understanding contrastive representation learning through alignment and uniformity on the hypersphere. In *Proceedings of the International Conference on Machine Learning*, pages 9929–9939, 2020.

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

[67] Yandong Wen, Kaipeng Zhang, Zhifeng Li, and Yu Qiao. A discriminative feature learning approach for deep face recognition. In *Proceedings of European Conference on Computer Vision*, pages 499–515, 2016.

[68] Aming Wu, Yahong Han, Linchao Zhu, and Yi Yang. Universal-prototype enhancing for few-shot object detection. In *Proceedings of IEEE/CVF International Conference on Computer Vision*, pages 9547–9556, 2021.

[69] Yongyu Xiong, Peipei Yang, and Cheng-Lin Liu. One-stage open set object detection with prototype learning. In *Proceedings of International Conference on Neural Information Processing*, pages 279–291, 2021.

[70] Hong-Ming Yang, Xu-Yao Zhang, Fei Yin, and Cheng-Lin Liu. Robust classification with convolutional prototype learning. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, pages 3474–3482, 2018.

[71] Jingkang Yang, Haoqi Wang, Litong Feng, Xiaopeng Yan, Huabin Zheng, Wayne Zhang, and Ziwei Liu. Semantically coherent out-of-distribution detection. In *Proceedings of the International Conference on Computer Vision*, pages 8281–8289, 2021.

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

[73] Fisher Yu, Haofeng Chen, Xin Wang, Wenqi Xian, Yingying Chen, Fangchen Liu, Vashisht Madhavan, and Trevor Darrell. BDD100K: A diverse driving dataset for heterogeneous multitask learning. In *Proceedings of IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2633–2642, 2020.

[74] Xuefei Zhe, Shifeng Chen, and Hong Yan. Directional statistics-based deep metric learning for image classification and retrieval. *Pattern Recognit.*, 93:113–123, 2019.

[75] Xizhou Zhu, Weijie Su, Lewei Lu, Bin Li, Xiaogang Wang, and Jifeng Dai. Deformable DETR: deformable transformers for end-to-end object detection. In *Proceedings of International Conference on Learning Representations*, 2021.

[76] Honglei Zhuang, Chi Wang, Fangbo Tao, Lance M. Kaplan, and Jiawei Han. Identifying semantically deviating outlier documents. In *Proceedings of Conference on Empirical Methods in Natural Language Processing*, pages 2748–2757, 2017.

Checklist

1. For all authors...
   (a) Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope? [Yes]
   (b) Have you read the ethics review guidelines and ensured that your paper conforms to them? [Yes]
   (c) Did you discuss any potential negative societal impacts of your work? [Yes] See Section 8.
   (d) Did you describe the limitations of your work? [N/A]
2. If you are including theoretical results...
   (a) Did you state the full set of assumptions of all theoretical results? [Yes]
   (b) Did you include complete proofs of all theoretical results? [Yes]
3. If you ran experiments...
   (a) Did you include the code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL)? [Yes] Code is released at `https://github.com/deeplearning-wisc/siren`.
   (b) Did you specify all the training details (e.g., data splits, hyperparameters, how they were chosen)? [Yes] See Section 4 and Appendix A and C.
   (c) Did you report error bars (e.g., with respect to the random seed after running experiments multiple times)? [Yes] See main results in Table 1.
   (d) Did you include the amount of compute and the type of resources used (e.g., type of GPUs, internal cluster, or cloud provider)? [Yes] See Appendix G.
4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets...
   (a) If your work uses existing assets, did you cite the creators? [Yes] See Section 4.
   (b) Did you mention the license of the assets? [N/A]
   (c) Did you include any new assets either in the supplemental material or as a URL? [N/A]
   (d) Did you discuss whether and how consent was obtained from people whose data you're using/curating? [N/A]
   (e) Did you discuss whether the data you are using/curating contains personally identifiable information or offensive content? [N/A]
5. If you used crowdsourcing or conducted research with human subjects...
   (a) Did you include the full text of instructions given to participants and screenshots, if applicable? [N/A]
   (b) Did you describe any potential participant risks, with links to Institutional Review Board (IRB) approvals, if applicable? [N/A]
   (c) Did you include the estimated hourly wage paid to participants and the total amount spent on participant compensation? [N/A]

SIREN: Shaping Representations for Detecting Out-of-distribution Objects (Appendix)

A Experimental Details

Following [13], we summarize the OOD detection evaluation task in Table 5. The OOD test dataset is selected from MS-COCO and OPENIMAGES dataset, which contains disjoint labels from the respective ID dataset. SIREN is trained for a total of 50 epochs on PASCAL-VOC, and trained for 30 epochs on BDD100K using ADAM optimizer. The initial learning rate is 2e-4 and decays at epoch 40 and 24 by 0.1 for PASCAL-VOC and BDD100K dataset, respectively. We set the number of the object queries as 300, the batch size as 8, and the weight β in Equation (8) as 1.5. See *detailed ablations on the hyperparameters on a validation OOD dataset in Appendix C*.

	Task 1	**Task 2**
ID train dataset	VOC train	BDD train
ID val dataset	VOC val	BDD val
OOD dataset	COCO & OPENIMAGES val	COCO and OPENIMAGES val
#ID train images	16,551	69,853
#ID val images	4,952	10,000
#OOD images for COCO	930	1,880
#OOD images for OPENIMAGES	1,761	1,761

Table 5: OOD detection evaluation tasks.

B Estimating $\widehat{\kappa}$

We provide the mathematical details for the estimation of κ in Equation (13). Concretely, we frame the estimation problem as deriving maximum likelihood estimators (MLEs) for the vMF density function, given the training data $\{\mathbf{r}_i\}_{i=1}^{M}$. Specifically, the maximum likelihood learning can be written as the following optimization problem:

$$\begin{aligned} &\max_{\boldsymbol{\mu},\kappa} \mathcal{L}_{\text{MLE}}(\boldsymbol{\mu},\kappa) := \sum_{i=1}^{M} \log p\left(\mathbf{r}_i\right) = M\kappa\boldsymbol{\mu}^\top\overline{\mathbf{r}} + M\log Z_d(\kappa) \\ &\text{s.t. } \|\boldsymbol{\mu}\| = 1 \text{ and } \kappa \geq 0, \end{aligned} \tag{14}$$

where $\overline{\mathbf{r}} = \frac{1}{M}\sum_{i=1}^{M}\mathbf{r}_i$ and $\mathcal{L}_{\text{MLE}}$ is the optimization objective of the maximum likelihood estimation for the distributional parameters $\boldsymbol{\mu},\kappa$. In order to optimize this objective, we take the derivatives of the objective *w.r.t.* the parameters $\boldsymbol{\mu},\kappa$ and set them to 0. Then, the optimal parameters $\boldsymbol{\mu}^*,\widehat{\kappa}$ should satisfy the following two conditions:

$$\boldsymbol{\mu}^* = \frac{\overline{\mathbf{r}}}{\|\overline{\mathbf{r}}\|}, \quad \frac{Z_d'\left(\widehat{\kappa}\right)}{Z_d\left(\widehat{\kappa}\right)} = -\|\overline{\mathbf{r}}\|, \tag{15}$$

where $Z_d'\left(\widehat{\kappa}\right) = \frac{dZ_d(\widehat{\kappa})}{d\widehat{\kappa}}$.

For the second condition in Equation (15), we let $\xi = (2\pi)^{d/2}$ and $s = d/2-1$ for notation simplicity. Put them into the Equation (2), we get the following:

$$Z_d(\widehat{\kappa}) = \frac{\widehat{\kappa}^s}{\xi\cdot I_s(\widehat{\kappa})}, \quad Z_d'(\widehat{\kappa}) = \frac{1}{\xi}\cdot\frac{s\widehat{\kappa}^{s-1}I_s(\widehat{\kappa}) - \widehat{\kappa}^s I_s'(\widehat{\kappa})}{I_s(\widehat{\kappa})^2}. \tag{16}$$

Divide $Z_d(\widehat{\kappa})$ by $Z_d'(\widehat{\kappa})$, we get:

$$\frac{Z_d'(\widehat{\kappa})}{Z_d(\widehat{\kappa})} = \frac{s}{\widehat{\kappa}} - \frac{I_s'(\widehat{\kappa})}{I_s(\widehat{\kappa})}. \tag{17}$$

Note that the Bessel function holds a recursive property, which is given as follows:

$$\frac{I_s'(\widehat{\kappa})}{I_s(\widehat{\kappa})} = \frac{s}{\widehat{\kappa}} + \frac{I_{s+1}'(\widehat{\kappa})}{I_s(\widehat{\kappa})}. \tag{18}$$

Substitute $\frac{I_s'(\widehat{\kappa})}{I_s(\widehat{\kappa})}$ in Equation (17) with the formula in Equation (18) and integrate it with the second condition of Equation (15), we get:

$$\frac{I_{d/2}(\widehat{\kappa})}{I_{d/2-1}(\widehat{\kappa})} = \|\overline{\mathbf{r}}\|. \tag{19}$$

Since there is no known closed-form solution to the Bessel ratio inversion problem as shown in the formula above, we adopt the approximation schemes based on the continued fraction form of the Bessel ratio function [1], namely:

$$R := \frac{I_{d/2}(\widehat{\kappa})}{I_{d/2-1}(\widehat{\kappa})} = \frac{1}{\frac{d}{\widehat{\kappa}} + \frac{1}{\frac{d+2}{\widehat{\kappa}} + \cdots}} \approx \frac{1}{\frac{d}{\widehat{\kappa}} + R}. \tag{20}$$

Replace R with $\|\overline{\mathbf{r}}\|$, we get $\widehat{\kappa} \approx \frac{d \cdot \|\overline{\mathbf{r}}\|}{1-\|\overline{\mathbf{r}}\|^2}$. Following [1], we further add a correction term $-\|\overline{\mathbf{r}}\|^3$ to the numerator and we get the approximated estimation calculated as:

$$\widehat{\kappa} = \frac{\|\overline{\mathbf{r}}\|(d - \|\overline{\mathbf{r}}\|^2)}{1 - \|\overline{\mathbf{r}}\|^2}. \tag{21}$$

C Hyperparameter Analysis

Below we perform sensitivity analysis for each important hyperparameter. Our sensitivity analysis uses the speckle-noised PASCAL-VOC validation dataset as OOD data, which is different from the actual OOD test datasets in use. We use SIREN pre-trained with DINO [5] as the object detection backbone, trained on in-distribution dataset PASCAL-VOC. We use the KNN score as the OOD score during inference.

Effect of the hypersphere dimension d. SIREN projects the object feature embeddings into a lower-dimensional hypersphere in $\mathbb{R}^d$, which allows tractable vMF estimation. A reasonable choice of the hyperspherical dimension d is able to preserve sufficient information for OOD detection while avoiding distributional parameter estimation in the high-dimension space. As shown in Table 6, using a dimension d between 16 and 64 yields a desirable and stable performance on the OOD validation data while properly maintaining the ID performance (mAP). We set $d = 16$ for PASCAL-VOC and 64 for BDD100K in Table 1.

d	mAP↑	FPR95 ↓	AUROC↑
8	60.4	76.31	64.95
16	**60.8**	**69.29**	**73.45**
32	58.1	75.53	70.32
64	58.7	74.55	69.47
80	58.2	69.09	72.29

Table 6: Ablation study on the dimension of the hypersphere d.

Effect of the SIREN loss weight β. In Table 7, we show the sensitivity of the OOD detection performance of our SIREN *w.r.t.* the weight β of the representation shaping loss. Overall, we find that $\beta = 1.5$ achieves the best OOD detection and ID performance.

Effect of the k **in the KNN score**. In Table 8, we show the sensitivity of the OOD detection performance of our SIREN *w.r.t.* the k of the KNN distance during inference. Overall, we find that $k = 10$ achieves the best OOD detection performance.

D Baselines

To evaluate the baselines, we follow the original methods in MSP [23], ODIN [35], KNN [60], and CSI [61] and apply them accordingly on the classification branch of the object detectors. The

β	mAP↑	FPR95 ↓	AUROC↑
0.1	60.2	69.72	73.21
0.5	59.2	71.09	72.36
1.0	59.8	76.20	70.49
1.5	**60.8**	**69.29**	**73.45**
2.0	58.9	70.41	71.22
2.5	56.0	77.20	68.37

Table 7: Ablation study on the loss weight β for $\mathcal{L}_{\text{SIREN}}$.

k	mAP↑	FPR95 ↓	AUROC↑
1	60.8	70.42	72.11
5	60.8	71.66	72.74
10	60.8	**69.29**	**73.45**
20	60.8	70.03	72.95
50	60.8	71.85	72.08
100	60.8	73.73	71.28
200	60.8	75.96	70.41

Table 8: Ablation study on the k for the KNN distance.

Mahalanobis distance [33] and gram matrices [53] are calculated based on the penultimate-layer features of the decoder in DDETR. For CSI [61], we use the rotation degree prediction (0°, 90°, 180°, 270°) as the self-supervised task. We set the temperature in the contrastive loss to 0.5. We use the penultimate-layer features of the decoder (with dimensionality 256) to perform contrastive learning. The weights of the losses that are used for classifying shifted instances and instance discrimination are both set to 1 following the original paper [61]. For OW-DETR [18], we follow the original paper and utilize the sigmoid probability of the additional unknown class for OOD detection. For VOS [13], we use the same hyperparameters as those in the original paper and synthesize virtual outliers in the penultimate layer of the object detector. Then we regard the virtual outliers as the negative samples during the object classification. For Dismax [39], we add the dismax loss with learnable prototypes in the penultimate layer of DDTER and apply the same inference score as the original paper for OOD detection.

E Comparison of Training Time

We provide the comparison of the training time for various baselines in Table 9 on our reported hardware (Section G in the Appendix). As the table shows, the training of our method SIREN incurs minimal computational overhead compared to the vanilla DDETR. In contrast, other baselines such as OW-DETR can be more than 2 times slower than SIREN.

Method	Training time (h)
ID: PASCAL-VOC / BDD100K	
Mahalanobis [33]	9.7 / 27
Gram matrices [53]	9.7 / 27
KNN [60]	9.7 / 27
CSI [61]	17.1 / 47.9
VOS [13]	11.4 / 32.7
OW-DETR [18]	23.7 / 59.3
Dismax [39]	10.0 / 27.7
SIREN (ours)	10.1 / 27.7

Table 9: Comparison of the training time for different baselines in Table 1 of the main paper.

F Details of Visualization

For Figure 3 in the main paper, we generate the toy data in the unit hypersphere by sampling from three vMF distributions in the 3D space. We adopt a concentration parameter κ of 100 for all three

classes. The centroid vectors are set to $[0, 0, 1]$, $[\frac{\sqrt{3}}{2}, 0, -\frac{1}{2}]$ and $[-\frac{\sqrt{3}}{2}, 0, -\frac{1}{2}]$, respectively. The uncertainty surface is obtained by calculating the uncertainty score of 200^2 points in the surface of the 3D ball.

G Software and Hardware

We run all experiments with Python 3.8.5 and PyTorch 1.7.0, using 8 NVIDIA GeForce RTX 2080Ti GPUs.