Out-of-Distribution Detection with Deep Nearest Neighbors

Yiyou Sun [1] Yifei Ming [1] Xiaojin Zhu [1] Yixuan Li [1]

Abstract

Out-of-distribution (OOD) detection is a critical task for deploying machine learning models in the open world. Distance-based methods have demonstrated promise, where testing samples are detected as OOD if they are relatively far away from in-distribution (ID) data. However, prior methods impose a strong distributional assumption of the underlying feature space, which may not always hold. In this paper, we explore the efficacy of non-parametric nearest-neighbor distance for OOD detection, which has been largely overlooked in the literature. Unlike prior works, our method does not impose any distributional assumption, hence providing stronger flexibility and generality. We demonstrate the effectiveness of nearest-neighbor-based OOD detection on several benchmarks and establish superior performance. Under the same model trained on ImageNet-1k, our method substantially reduces the false positive rate (FPR@TPR95) by 24.77% compared to a strong baseline SSD+, which uses a parametric approach Mahalanobis distance in detection. Code is available: https://github.com/deeplearning-wisc/knn-ood.

1. Introduction

Modern machine learning models deployed in the open world often struggle with out-of-distribution (OOD) inputs—samples from a different distribution that the network has not been exposed to during training, and therefore should not be predicted at test time. A reliable classifier should not only accurately classify known in-distribution (ID) samples, but also identify as "unknown" any OOD input. This gives rise to the importance of OOD detection, which determines whether an input is ID or OOD and enables the model to take precautions.

A rich line of OOD detection algorithms has been developed recently, among which distance-based methods demonstrated promise (Lee et al., 2018; Tack et al., 2020; Sehwag et al., 2021). Distance-based methods leverage feature embeddings extracted from a model, and operate under the assumption that the test OOD samples are relatively far away from the ID data. For example, Lee et al. modeled the feature embedding space as a mixture of multivariate Gaussian distributions, and used the maximum Mahalanobis distance (Mahalanobis, 1936) to all class centroids for OOD detection. However, all these approaches make a strong distributional assumption of the underlying feature space being class-conditional Gaussian. As we verify, the learned embeddings can fail the Henze-Zirkler multivariate normality test (Henze & Zirkler, 1990). This limitation leads to the open question:

Can we leverage the non-parametric nearest neighbor approach for OOD detection?

Unlike prior works, the non-parametric approach does not impose any distributional assumption about the underlying feature space, hence providing stronger *flexibility and generality*. Despite its simplicity, the nearest neighbor approach has received scant attention. Looking at the literature on OOD detection in the past several years, there has not been any work that demonstrated the efficacy of a non-parametric nearest neighbor approach for this problem. This suggests that making the seemingly simple idea work is non-trivial. Indeed, we found that simply using the nearest neighbor distance derived from the feature embedding of a standard classification model is not performant.

In this paper, we challenge the status quo by presenting the first study exploring and demonstrating the efficacy of the non-parametric nearest-neighbor distance for OOD detection. To detect OOD samples, we compute the k-th nearest neighbor (KNN) distance between the embedding of test input and the embeddings of the training set and use a threshold-based criterion to determine if the input is OOD or not. In a nutshell, we perform non-parametric level set estimation, partitioning the data into two sets (ID vs. OOD) based on the deep k-nearest neighbor distance. KNN offers compelling advantages of being: (1) **distributional assumption free**, (2) **OOD-agnostic** (*i.e.*, the distance threshold is estimated on the ID data only, and does not rely on

[1] Department of Computer Sciences, University of Wisconsin - Madison. Correspondence to: Yiyou Sun, Yixuan Li <sunyiyou, sharonli@cs.wisc.edu>.

Proceedings of the 39th International Conference on Machine Learning, Baltimore, Maryland, USA, PMLR 162, 2022. Copyright 2022 by the author(s).

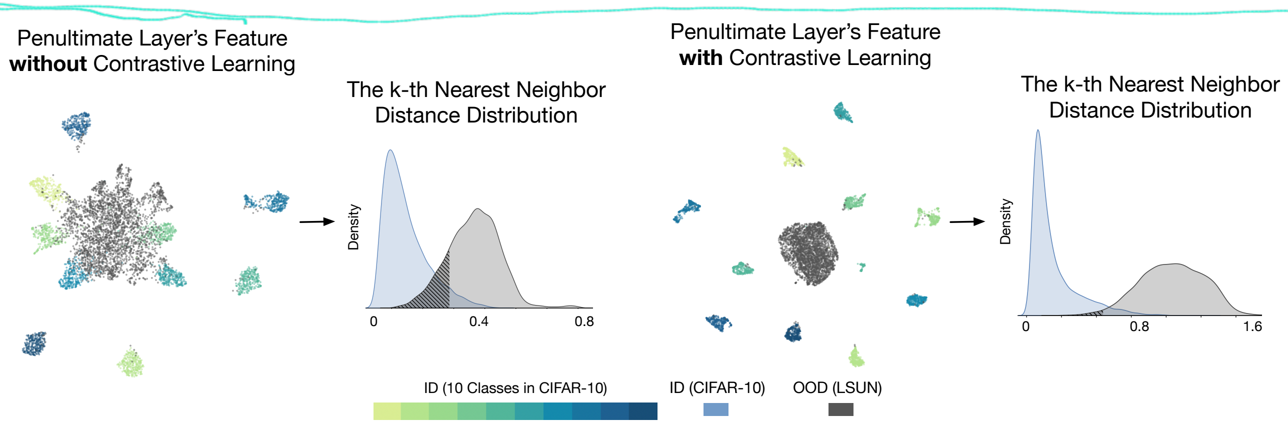

Figure 1. Illustration of our framework using nearest neighbors for OOD detection. KNNperforms non-parametric level set estimation, partitioning the data into two sets (ID vs. OOD) based on the k-th nearest neighbor distance. The distances are estimated from the penultimate feature embeddings, visualized via UMAP (McInnes et al., 2018). Models are trained on ResNet-18 (He et al., 2016) using cross-entropy loss (left) v.s. contrastive loss (right). The in-distribution data is CIFAR-10 (colored in non-gray colors) and OOD data is LSUN (colored in gray). The shaded grey area in the density distribution plot indicates OOD samples that are misidentified as ID data.

information of unknown data), (3) **easy-to-use** (*i.e.*, no need to calculate the inverse of the covariance matrix which can be numerically unstable), and (4) **model-agnostic** (*i.e.*, the testing procedure is applicable to different model architectures and training losses).

Our exploration leads to both empirical effectiveness (Section 4 & 5) and theoretical justification (Section 6). By studying the role of representation space, we show that a compact and normalized feature space is the key to the success of the nearest neighbor approach for OOD detection. Extensive experiments show that KNN outperforms the parametric approach, and scales well to the large-scale dataset. Computationally, modern implementations of approximate nearest neighbor search allow us to do this in milliseconds even when the database contains billions of images (Johnson et al., 2019). On a challenging ImageNet OOD detection benchmark (Huang & Li, 2021), our KNN-based approach achieves superior performance under a similar inference speed as the baseline methods. The overall simplicity and effectiveness of KNN make it appealing for real-world applications. We summarize our contributions below:

1. We present the first study exploring and demonstrating the efficacy of non-parametric density estimation with nearest neighbors for OOD detection—a simple, flexible yet overlooked approach in literature. We hope our work draws attention to the strong promise of the non-parametric approach, which obviates data assumption on the feature space.

2. We demonstrate the superior performance of the KNN-based method on several OOD detection benchmarks, different model architectures (including CNNs and ViTs), and different training losses. Under the same model trained on ImageNet-1k, our method substantially reduces the false positive rate (FPR@TPR95) by **24.77%** compared to a strong baseline SSD+ (Sehwag et al., 2021), which uses a parametric approach (*i.e.*, Mahalanobis distance (Lee et al., 2018)) for detection.

3. We offer new insights on the key components to make KNN effective in practice, including feature normalization and a compact representation space. Our findings are supported by extensive ablations and experiments. We believe these insights are valuable to the community in carrying out future research.

4. We provide theoretical analysis, showing that KNN-based OOD detection can reject inputs equivalent to the Bayes optimal estimator. By modeling the nearest neighbor distance in the feature space, our theory (1) directly connects to our method which also operates in the feature space, and (2) complements our experiments by considering the universality of OOD data.

2. Preliminaries

We consider supervised multi-class classification, where $\mathcal{X}$ denotes the input space and $\mathcal{Y} = \{1, 2, ..., C\}$ denotes the label space. The training set $\mathbb{D}_{in} = \{(\mathbf{x}_i, y_i)\}_{i=1}^n$ is drawn *i.i.d.* from the joint data distribution $P_{\mathcal{X}\mathcal{Y}}$. Let $\mathcal{P}_{\text{in}}$ denote the marginal distribution on $\mathcal{X}$. Let $f : \mathcal{X} \mapsto \mathbb{R}^{|\mathcal{Y}|}$ be a neural network trained on samples drawn from $P_{\mathcal{X}\mathcal{Y}}$ to output a logit vector, which is used to predict the label of an input sample.

Out-of-distribution detection When deploying a machine model in the real world, a reliable classifier should not only accurately classify known in-distribution (ID) samples, but also identify as "unknown" any OOD input. This can be achieved by having an OOD detector, in tandem with the classification model f.

OOD detection can be formulated as a binary classification problem. At test time, the goal of OOD detection is to decide whether a sample $\mathbf{x} \in \mathcal{X}$ is from $\mathcal{P}_{\text{in}}$ (ID) or not (OOD). The decision can be made via a level set estimation:

$$G_\lambda(x) = \begin{cases} \text{ID} & S(\mathbf{x}) \geq \lambda \\ \text{OOD} & S(\mathbf{x}) < \lambda \end{cases},$$

where samples with higher scores $S(\mathbf{x})$ are classified as ID and vice versa, and λ is the threshold. In practice, OOD is often defined by a distribution that simulates unknowns encountered during deployment time, such as samples from an irrelevant distribution whose label set has no intersection with $\mathcal{Y}$ and therefore should not be predicted by the model.

3. Deep Nearest Neighbor for OOD detection

In this section, we describe our approach using the deep k-Nearest Neighbor (KNN) for OOD detection. We illustrate our approach in Figure 1, which at a high level, can be categorized as a distance-based method. Distance-based methods leverage feature embeddings extracted from a model and operate under the assumption that the test OOD samples are relatively far away from the ID data. Previous distance-based OOD detection methods employed parametric density estimation and modeled the feature embedding space as a mixture of multivariate Gaussian distributions (Lee et al., 2018). However, such an approach makes a strong distributional assumption of the learned feature space, which may not necessarily hold[1].

In this paper, we instead explore the efficacy of *non-parametric density estimation using nearest neighbors* for OOD detection. Despite the simplicity, KNN approach is not systematically explored or compared in most current OOD detection papers. Specifically, we compute the k-th nearest neighbor distance between the embedding of each test image and the training set, and use a simple threshold-based criterion to determine if an input is OOD or not. Importantly, we use the normalized penultimate feature $\mathbf{z} = \phi(\mathbf{x})/\|\phi(\mathbf{x})\|_2$ for OOD detection, where $\phi : \mathcal{X} \mapsto \mathbb{R}^m$ is a feature encoder. Denote the embedding set of training data as $\mathbb{Z}_n = (\mathbf{z}_1, \mathbf{z}_2, ..., \mathbf{z}_n)$. During testing, we derive the normalized feature vector $\mathbf{z}^*$ for a test sample $\mathbf{x}^*$, and calculate the Euclidean distances $\|\mathbf{z}_i - \mathbf{z}^*\|_2$ with respect to embedding vectors $\mathbf{z}_i \in \mathbb{Z}_n$. We reorder $\mathbb{Z}_n$ according to the increasing distance $\|\mathbf{z}_i - \mathbf{z}^*\|_2$. Denote the reordered data sequence as $\mathbb{Z}'_n = (\mathbf{z}_{(1)}, \mathbf{z}_{(2)}, ..., \mathbf{z}_{(n)})$. The decision function for OOD detection is given by:

$$G(\mathbf{z}^*; k) = \mathbf{1}\{-r_k(\mathbf{z}^*) \geq \lambda\},$$

where $r_k(\mathbf{z}^*) = \|\mathbf{z}^* - \mathbf{z}_{(k)}\|_2$ is the distance to the k-th nearest neighbor (k-NN) and $\mathbf{1}\{\cdot\}$ is the indicator function. The threshold λ is typically chosen so that a high fraction of ID data (*e.g.,* 95%) is correctly classified. The threshold does not depend on OOD data.

Algorithm 1 OOD Detection with Deep Nearest Neighbors

```
Input: Training dataset $\mathbb{D}_{in}$, pre-trained neural network encoder $\phi$, test sample $\mathbf{x}^*$, threshold $\lambda$
For $\mathbf{x}_i$ in the training data $\mathbb{D}_{in}$, collect feature vectors $\mathbb{Z}_n = (\mathbf{z}_1, \mathbf{z}_2, ..., \mathbf{z}_n)$
Testing Stage:
Given a test sample, we calculate feature vector $\mathbf{z}^* = \phi(\mathbf{x}^*)/\|\phi(\mathbf{x}^*)\|_2$
Reorder $\mathbb{Z}_n$ according to the increasing value of $\|\mathbf{z}_i - \mathbf{z}^*\|_2$ as $\mathbb{Z}'_n = (\mathbf{z}_{(1)}, \mathbf{z}_{(2)}, ..., \mathbf{z}_{(n)})$
Output: OOD detection decision $\mathbf{1}\{-\|\mathbf{z}^* - \mathbf{z}_{(k)}\|_2 \geq \lambda\}$
```

We summarize our approach in Algorithm 1. Noticeably, KNN-based OOD detection offers several compelling advantages:

1. **Distributional assumption free**: Non-parametric nearest neighbor approach does not impose distributional assumptions about the underlying feature space. KNN therefore provides stronger flexibility and generality, and is applicable even when the feature space does not conform to the mixture of Gaussians.

2. **OOD-agnostic**: The testing procedure does not rely on the information of unknown data. The distance threshold is estimated on the ID data only.

3. **Easy-to-use**: Modern implementations of approximate nearest neighbor search allow us to do this in milliseconds even when the database contains billions of images (Johnson et al., 2019). In contrast, Mahalanobis distance requires calculating the inverse of the covariance matrix, which can be numerically unstable.

4. **Model-agnostic**: The testing procedure applies to a variety of model architectures, including CNNs and more recent Transformer-based ViT models (Dosovitskiy et al., 2021). Moreover, we will show that KNN is agnostic to the training procedure as well, and is compatible with models trained under different loss functions (*e.g.*, cross-entropy loss and contrastive loss).

We proceed to show the efficacy of the KNN-based OOD detection approach in Section 4.

[1] We verified this by performing the Henze-Zirkler multivariate normality test (Henze & Zirkler, 1990) on the embeddings. The testing results show that the feature vectors for each class are not normally distributed at the significance level of 0.05.

Table 1. **Results on CIFAR-10.** Comparison with competitive OOD detection methods. All methods are based on a discriminative model trained on ID data only, without using outlier data. ↑ indicates larger values are better and vice versa.

Method	OOD Dataset												ID ACC
	SVHN		LSUN		iSUN		Texture		Places365		Average		
	FPR↓	AUROC↑	FPR↓	AUROC↑	FPR↓	AUROC↑	FPR↓	AUROC↑	FPR↓	AUROC↑	FPR↓	AUROC↑	
Without Contrastive Learning													
MSP	59.66	91.25	45.21	93.80	54.57	92.12	66.45	88.50	62.46	88.64	57.67	90.86	94.21
ODIN	53.78	91.30	10.93	97.93	28.44	95.51	55.59	89.47	43.40	90.98	38.43	93.04	94.21
Energy	54.41	91.22	10.19	98.05	27.52	95.59	55.23	89.37	42.77	91.02	38.02	93.05	94.21
GODIN	18.72	96.10	11.52	97.12	30.02	94.02	33.58	92.20	55.25	85.50	29.82	92.97	93.64
Mahalanobis	9.24	97.80	67.73	73.61	6.02	98.63	23.21	92.91	83.50	69.56	37.94	86.50	94.21
KNN (ours)	27.97	95.48	18.50	96.84	24.68	95.52	26.74	94.96	47.84	89.93	29.15	94.55	94.21
With Contrastive Learning													
CSI	37.38	94.69	5.88	98.86	10.36	98.01	28.85	94.87	38.31	93.04	24.16	95.89	94.38
SSD+	1.51	99.68	6.09	98.48	33.60	95.16	12.98	97.70	28.41	94.72	16.52	97.15	**95.07**
KNN+ (ours)	2.42	99.52	1.78	99.48	20.06	96.74	8.09	98.56	23.02	95.36	**11.07**	**97.93**	**95.07**

4. Experiments

The goal of our experimental evaluation is to answer the following questions: (1) How does KNN fare against the parametric counterpart such as Mahalanobis distance for OOD detection? (2) Can KNN scale to a more challenging task when the training data is large-scale (*e.g.*, ImageNet)? (3) Is KNN-based OOD detection effective under different model architectures and training objectives? (4) How do various design choices affect the performance?

Evaluation metrics We report the following metrics: (1) the false positive rate (FPR95) of OOD samples when the true positive rate of ID samples is at 95%, (2) the area under the receiver operating characteristic curve (AUROC), (3) ID classification accuracy (ID ACC), and (4) per-image inference time (in milliseconds, averaged across test images).

Training losses In our experiments, we aim to show that KNN-based OOD detection is agnostic to the training procedure, and is compatible with models trained under different losses. We consider two types of loss functions, with and without contrastive learning respectively. We employ (1) cross-entropy loss which is the most commonly used training objective in classification, and (2) supervised contrastive learning (SupCon) (Khosla et al., 2020)— the latest development for representation learning, which leverages the label information by aligning samples belonging to the same class in the embedding space.

Remark on the implementation All of the experiments are based on PyTorch (Paszke et al., 2019). Code is made publicly available online. We use Faiss (Johnson et al., 2019), a library for efficient nearest neighbor search. Specifically, we use faiss.IndexFlatL2 as the indexing method with Euclidean distance. In practice, we precompute the embeddings for all images and store them in a key-value map to make KNN search efficient. The embedding vectors for ID data only need to be extracted once after the training is completed.

4.1. Evaluation on Common Benchmarks

Datasets We begin with the CIFAR benchmarks that are routinely used in literature. We use the standard split with 50,000 training images and 10,000 test images. We evaluate the methods on common OOD datasets: Textures (Cimpoi et al., 2014), SVHN (Netzer et al., 2011), Places365 (Zhou et al., 2017), LSUN-C (Yu et al., 2015), and iSUN (Xu et al., 2015). All images are of size 32×32.

Experiment details We use ResNet-18 as the backbone for CIFAR-10. Following the original settings in Khosla et al., models with SupCon loss are trained for 500 epochs, with the batch size of 1024. The temperature τ is 0.1. The dimension of the penultimate feature where we perform the nearest neighbor search is 512. The dimension of the projection head is 128. We use the cosine annealing learning rate (Loshchilov & Hutter, 2016) starting at 0.5. We use $k = 50$ for CIFAR-10 and $k = 200$ for CIFAR-100, which is selected from $k = \{1, 10, 20, 50, 100, 200, 500, 1000, 3000, 5000\}$ using the validation method in (Hendrycks et al., 2019). We train the models using stochastic gradient descent with momentum 0.9, and weight decay 10^{-4}. The model without contrastive learning is trained for 100 epochs. The start learning rate is 0.1 and decays by a factor of 10 at epochs 50, 75, and 90 respectively.

Nearest neighbor distance achieves superior performance We present results in Table 1, where non-parametric KNN approach shows favorable performance. Our comparison covers an extensive collection of competitive methods in the literature. For clarity, we divide the baseline methods into two categories: trained with and without contrastive losses. Several baselines derive OOD scores from a model trained with common softmax cross-entropy (CE) loss, including MSP (Hendrycks & Gimpel, 2017), ODIN (Liang et al., 2018), Mahalanobis (Lee

Table 2. Evaluation (FPR95) on hard OOD detection tasks. Model is trained on CIFAR-10 with SupCon loss.

	LSUN-FIX	ImageNet-FIX	ImageNet-R	C-100
SSD+	29.86	32.26	45.62	45.50
KNN+ (Ours)	**21.52**	**25.92**	**29.92**	**38.83**

et al., 2018), and Energy (Liu et al., 2020). GODIN (Hsu et al., 2020) is trained using a DeConf-C loss, which does not involve contrastive loss either. For methods involving contrastive losses, we use the same network backbone architecture and embedding dimension, while only varying the training objective. These methods include CSI (Tack et al., 2020) and SSD+ (Sehwag et al., 2021). For terminology clarity, KNN refers to our method trained with CE loss, and KNN+ refers to the variant trained with SupCon loss. We highlight two groups of comparisons:

- **KNN vs. Mahalanobis** (without contrastive learning): Under the *same* model trained with cross-entropy (CE) loss, our method achieves an average FPR95 of 29.15%, compared to that of Mahalanobis distance 37.94%. The performance gain precisely demonstrates the advantage of KNN over the parametric method Mahalanobis distance.
- **KNN+ vs. SSD+** (with contrastive loss): KNN+ and SSD+ are fundamentally different in OOD detection mechanisms, despite both benefit from the contrastively learned representations. SSD+ modeled the feature embedding space as a multivariate Gaussian distribution for each class, and use Mahalanobis distance (Lee et al., 2018) for OOD detection. Under the *same* model trained with Supervised Contrastive Learning (SupCon) loss, our method with the nearest neighbor distance reduces the average FPR95 by 5.45%, which is a relatively **32.99%** reduction in error. It further suggests the advantage of using nearest neighbors without making any distributional assumptions on the feature embedding space.

The above comparison suggests that the nearest neighbor approach is compatible with models trained both with and without contrastive learning. In addition, KNN is also simpler to use and implement than CSI, which relies on sophisticated data augmentations and ensembling in testing. Lastly, as a result of the improved embedding quality, the ID accuracy of the model trained with SupCon loss is improved by 0.86% on CIFAR-10 and 2.45% on ImageNet compared to training with the CE loss. Due to space constraints, we provide results on DenseNet (Huang et al., 2017) in Appendix C.

Contrastively learned representation helps While contrastive learning has been extensively studied in recent literature, its role remains untapped when coupled with a non-parametric approach (such as nearest neighbors) for OOD detection. We examine the effect of using supervised contrastive loss for KNN-based OOD detection. We provide both qualitative and quantitative evidence, highlighting advantages over the standard softmax cross-entropy (CE) loss. (1) We visualize the learned feature embeddings in Figure 1 using UMAP (McInnes et al., 2018), where the colors encode different class labels. A salient observation is that the representation with SupCon is more distinguishable and compact than the representation obtained from the CE loss. The high-quality embedding space indeed confers benefits for KNN-based OOD detection. (2) Beyond visualization, we also quantitatively compare the performance of KNN-based OOD detection using embeddings trained with SupCon vs CE. As shown in Table 1, KNN+ with contrastively learned representations reduces the FPR95 on all test OOD datasets compared to using embeddings from the model trained with CE loss.

Comparison with other non-parametric methods In Table 3, we compare the nearest neighbor approach with other non-parametric methods. For a fair comparison, we use the same embeddings trained with SupCon loss. Our comparison covers an extensive collection of outlier detection methods in literature including: IForest (Liu et al., 2008), OCSVM (Schölkopf et al., 2001), LODA (Pevnỳ, 2016), PCA (Shyu et al., 2003), and LOF (Breunig et al., 2000). The parameter setting for these methods is available in Appendix B. We show that KNN+ outperforms alternative non-parametric methods by a large margin.

Table 3. Comparison with other non-parametric methods. Results are averaged across all test OOD datasets. Model is trained on CIFAR-10.

	FPR95↓	AUROC↑
IForest (Liu et al., 2008)	65.49	76.98
OCSVM (Schölkopf et al., 2001)	52.27	65.16
LODA (Pevnỳ, 2016)	76.38	62.59
PCA (Shyu et al., 2003)	37.26	83.13
LOF (Breunig et al., 2000)	40.06	93.47
KNN+ (ours)	**11.07**	**97.93**

Evaluations on hard OOD tasks Hard OOD samples are particularly challenging to detect. To test the limit of the non-parametric KNN approach, we follow CSI (Tack et al., 2020) and evaluate on several hard OOD datasets: LSUN-

Table 4. **Results on ImageNet**. All methods are based on a model trained on ID data only (ImageNet-1k (Deng et al., 2009)). We report the OOD detection performance, along with the per-image inference time.

Methods	Inference time (ms)	iNaturalist FPR95 ↓	iNaturalist AUROC ↑	SUN FPR95 ↓	SUN AUROC ↑	Places FPR95 ↓	Places AUROC ↑	Textures FPR95 ↓	Textures AUROC ↑	Average FPR95 ↓	Average AUROC ↑	ID ACC
Without Contrastive Learning												
MSP	7.04	54.99	87.74	70.83	80.86	73.99	79.76	68.00	79.61	66.95	81.99	75.08
ODIN	7.05	47.66	89.66	60.15	84.59	67.89	81.78	50.23	85.62	56.48	85.41	75.08
Energy	7.04	55.72	89.95	59.26	85.89	64.92	82.86	53.72	85.99	58.41	86.17	75.08
GODIN	7.04	61.91	85.40	60.83	85.60	63.70	83.81	77.85	73.27	66.07	82.02	70.43
Mahalanobis	35.83	97.00	52.65	98.50	42.41	98.40	41.79	55.80	85.01	87.43	55.47	75.08
KNN ($\alpha = 100\%$)	10.31	59.77	85.89	68.88	80.08	78.15	74.10	10.90	97.42	54.68	84.37	75.08
KNN ($\alpha = 1\%$)	7.04	59.08	86.20	69.53	80.10	77.09	74.87	11.56	97.18	54.32	84.59	75.08
With Contrastive Learning												
SSD+	28.31	57.16	87.77	78.23	73.10	81.19	70.97	36.37	88.52	63.24	80.09	**79.10**
KNN+ ($\alpha = 100\%$)	10.47	30.18	94.89	48.99	88.63	59.15	84.71	15.55	95.40	**38.47**	**90.91**	**79.10**
KNN+ ($\alpha = 1\%$)	7.04	30.83	94.72	48.91	88.40	60.02	84.62	16.97	94.45	39.18	90.55	**79.10**

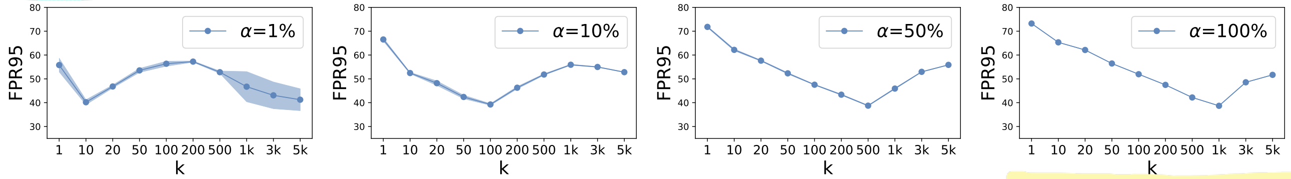

Figure 2. Comparison with the effect of different k and sampling ratio α. We report an average FPR95 score over four test OOD datasets. The variances are estimated across 5 different random seeds. The solid blue line represents the averaged value across all runs and the shaded blue area represents the standard deviation. Note that the full ImageNet dataset ($\alpha = 100\%$) has 1000 images per class.

FIX, ImageNet-FIX, ImageNet-R, and CIFAR-100. The results are summarized in Table 2. Under the same model, KNN+ consistently outperforms SSD+.

4.2. Evaluation on Large-scale ImageNet Task

We evaluate on a large-scale OOD detection task based on ImageNet (Deng et al., 2009). Compared to the CIFAR benchmarks above, the ImageNet task is more challenging due to a large amount of training data. Our goal is to verify KNN's performance benefits and whether it scales computationally with millions of samples.

Setup We use a ResNet-50 backbone (He et al., 2016) and train on ImageNet-1k (Deng et al., 2009) with resolution 224 × 224. Following the experiments in Khosla et al., models with SupCon loss are trained for 700 epochs, with a batch size of 1024. The temperature τ is 0.1. The dimension of the penultimate feature where we perform the nearest neighbor search is 2048. The dimension of the project head is 128. We use the cosine learning rate (Loshchilov & Hutter, 2016) starting at 0.5. We train the models using stochastic gradient descent with momentum 0.9, and weight decay 10^{-4}. We use $k = 1000$ which follows the same validation procedure as before. When randomly sampling $\alpha\%$ training data for nearest neighbor search, k is scaled accordingly to $1000 \cdot \alpha\%$.

Following the ImageNet-based OOD detection benchmark in MOS (Huang & Li, 2021), we evaluate on four test OOD datasets that are subsets of: Places365 (Zhou et al., 2017), Textures (Cimpoi et al., 2014), iNaturalist (Van Horn et al., 2018), and SUN (Xiao et al., 2010) with non-overlapping categories *w.r.t.* ImageNet. The evaluations span a diverse range of domains including fine-grained images, scene images, and textural images.

Nearest neighbor approach achieves superior performance without compromising the inference speed In Table 4, we compare our approach with OOD detection methods that are competitive in the literature. The baselines are the same as what we described in Section 4.1 except for CSI[2]. We report both OOD detection performance and the inference time (measured by milliseconds). We highlight three trends: (1) KNN+ outperforms the best baseline by **18.01**% in FPR95. (2) Compared to SSD+, KNN+ substantially reduces the FPR95 by **24.77**% averaged across all test sets. The limiting performance of SSD+ is due to the increased size of label space and data complexity, which makes the class-conditional Gaussian assumption less viable. In contrast, our non-parametric method does not suffer from this issue, and can better estimate the density of the com-

[2] The training procedure of CSI is computationally prohibitive on ImageNet, which takes three months on 8 Nvidia 2080Tis.

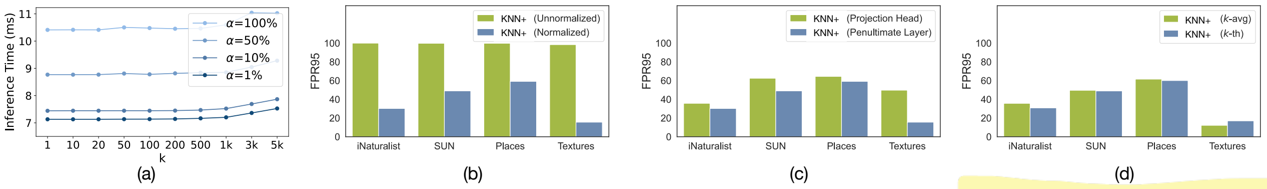

Figure 3. **Ablation results.** In (a), we compare the inference speed (per-image) using different k and sampling ration α. For (b) (c) (d), the FPR95 value is reported over all test OOD datasets. Specifically, (b) compares the effect of using normalization in the penultimate layer feature vs. without normalization, (c) compares using features in the penultimate layer feature vs the projection head, and (d) compares the OOD detection performance using k-th and averaged k (k-avg) nearest neighbor distance.

plex distribution for OOD detection. (3) KNN+ achieves strong performance with a comparable inference speed as the baselines. In particular, we show that performing nearest neighbor distance estimation with only 1% randomly sampled training data can yield a similar performance as using the full dataset.

Nearest neighbor approach is competitive on ViT Going beyond convolutional neural networks, we show in Table 5 that the nearest neighbor approach is effective for transformer-based ViT model (Dosovitskiy et al., 2021). We adopt the ViT-B/16 architecture fine-tuned on the ImageNet-1k dataset using cross-entropy loss. Under the same ViT model, our non-parametric KNN method consistently outperforms Mahalanobis.

5. A Closer Look at KNN-based OOD Detection

We provide further analysis and ablations to understand the behavior of KNN-based OOD detection. All the ablations are based on the ImageNet model trained with SupCon loss (same as in Section 4.2).

Effect of k and sampling ratio In Figure 2 and Figure 3 (a), we systematically analyze the effect of k and the dataset sampling ratios α. We vary the number of neighbors $k = \{1, 10, 20, 50, 100, 200, 500, 1000, 3000, 5000\}$ and random sampling ratio $\alpha = \{1\%, 10\%, 50\%, 100\%\}$. We note several interesting observations: (1) The optimal OOD detection (measured by FPR95) remains *similar* under different random sampling ratios α. (2) The optimal k is consistent with the one chosen by our validation strategy. For example, the optimal k is 1,000 when $\alpha = 100\%$; and the optimal k becomes 10 when $\alpha = 1\%$. (3) Varying k does not significantly affect the inference speed when k is relatively small (*e.g.*, $k < 1000$) as shown in Figure 3 (a).

Feature normalization is critical In this ablation, we contrast the performance of KNN-based OOD detection with and without feature normalization. The k-th NN distance can be derived by $r_k(\frac{\phi(\mathbf{x})}{\|(\phi(\mathbf{x})\|})$ and $r_k(\phi(\mathbf{x}))$, respectively. As shown in Figure 3 (b), using feature normalization improved the FPR95 drastically by **61.05%**, compared to the counterpart without normalization. To better understand this, we look into the Euclidean distance $r = \|u - v\|_2$ between two vectors u and v. The norm of the feature vector u and v could notably affect the value of the Euclidean distance. Interestingly, recent studies share the observation in Figure 4 (a) that the ID data has a larger L_2 feature norm than OOD data (Tack et al., 2020; Huang et al., 2021). Therefore, the Euclidean distance between ID features can be large (Figure 4 (b)). This contradicts the hope that ID data has a smaller k-NN distance than OOD data. Indeed, the normalization effectively mitigated this problem, as evidenced in Figure 4 (c). Empirically, the normalization plays a key role in the nearest neighbor approach to be successful in OOD detection as shown in Figure 3 (b).

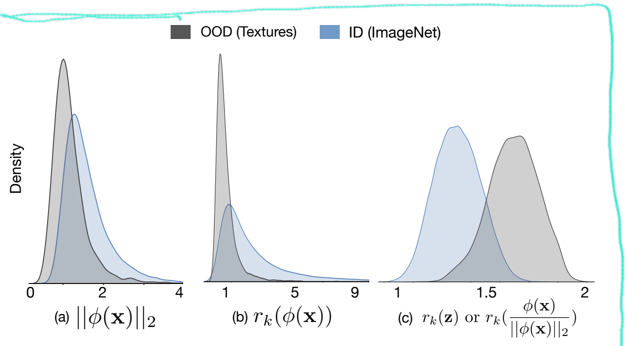

Figure 4. Distribution of (a) the L_2-norm of feature embeddings, (b) the k-NN distance with the *unnormalized* feature embeddings, and (c) the k-NN distance with the *normalized* features.

Using the penultimate layer's feature is better than using the projection head In this paper, we follow the convention in `SSD+`, which uses features from the penultimate layer instead of the projection head. We also verify in Figure 3 (c) that using the penultimate layer's feature is better than using the projection head on all test OOD datasets. This is likely due to the penultimate layer preserves more information than the projection head, which has much smaller dimensions.

Table 5. Performance comparison (FPR95) on ViT-B/16 model fine-tuned on ImageNet-1k.

	iNaturalist	**SUN**	**Places**	**Textures**
Mahalanobis (parametric)	17.56	80.51	84.12	70.51
KNN (non-parametric)	**7.30**	**48.40**	**56.46**	**39.91**

KNN can be further boosted by activation rectification We show that KNN+ can be made stronger with a recent method of activation rectification (Sun et al., 2021). It was shown that the OOD data can have overly high activations on some feature dimensions, and this rectification is effective in suppressing the values. Empirically, we compare the results in Table 6 by using the activation rectification and achieve improved OOD detection performance.

Table 6. Comparison of KNN-based method with and without activation truncation. The ID data is ImageNet-1k. The value is averaged over all test OOD datasets.

Method	FPR95↓	AUROC↑
KNN+	38.47	90.91
KNN+ (w. ReAct (Sun et al., 2021))	**26.45**	**93.76**

Using k-th and averaged k nearest nerighbors' distance has similar performance We compare two variants for OOD detection: k-th nearest neighbor distance vs. averaged k (k-avg) nearest neighbor distance. The comparison is shown in Figure 3 (d), where the average performance (on four datasets) is on par. The reported results are based on the full ID dataset ($\alpha = 100\%$) with the optimal k chosen for k-th NN and k-avg NN respectively. Despite the similar performance, using k-th NN distance has a stronger theoretical interpretation, as we show in the next section.

6. Theoretical Justification

In this section, we provide a theoretical analysis of using KNN for OOD detection. By modeling the KNN in the feature space, our theory (1) directly connects to our method which also operates in the feature space, and (2) complements our experiments by considering the universality of OOD data. Our goal here is to analyze the average performance of our algorithm while being OOD-agnostic and training-agnostic.

Setup We consider OOD detection task as a special binary classification task, where the negative samples (OOD) are only available in the testing stage. We assume the input is from feature embeddings space $\mathcal{Z}$ and the labeling set $\mathcal{G} = \{0(\text{OOD}), 1(\text{ID})\}$. In the inference stage, the testing set $\{(\mathbf{z}_i, g_i)\}$ is drawn *i.i.d.* from $P_{\mathcal{Z}\mathcal{G}}$.

Denote the marginal distribution on $\mathcal{Z}$ as $\mathcal{P}$. We adopt the Huber contamination model (Huber, 1964) to model the fact that we may encounter both ID and OOD data in test time:

$$\mathcal{P} = \varepsilon \mathcal{P}_{out} + (1-\varepsilon)\mathcal{P}_{in},$$

where $\mathcal{P}_{in}$ and $\mathcal{P}_{out}$ are the underlying distributions of feature embeddings for ID and OOD data, respectively, and ε is a constant controlling the fraction of OOD samples in testing. We use lower case $p_{in}(\mathbf{z}_i)$ and $p_{out}(\mathbf{z}_i)$ to denote the probability density function, where $p_{in}(\mathbf{z}_i) = p(\mathbf{z}_i|g_i = 1)$ and $p_{out}(\mathbf{z}_i) = p(\mathbf{z}_i|g_i = 0)$.

A key challenge in OOD detection (and theoretical analysis) is the lack of knowledge on OOD distribution, which can arise universally outside ID data. We thus try to keep our analysis general and reflect the fact that we do not have any strong prior information about OOD. For this reason, we model OOD data with an equal chance to appear outside of the high-density region of ID data, $p_{out}(\mathbf{z}) = c_0 \mathbf{1}\{p_{in}(\mathbf{z}) < c_1\}$[3]. The Bayesian classifier is known as the optimal binary classifier defined by $h_{Bay}(\mathbf{z}_i) = \mathbf{1}\{p(g_i = 1|\mathbf{z}_i) \geq \beta\}$[4], assuming the underlying density function is given.

Without such oracle information, our method applies k-NN as the distance measure which acts as a probability density estimation, and thus provides the decision boundary based on it. Specifically, KNN's hypothesis class $\mathcal{H}$ is given by $\{h : h_{\lambda,k,\mathbb{Z}_n}(\mathbf{z}_i) = \mathbf{1}\{-r_k(\mathbf{z}_i) \geq \lambda\}\}$, where $r_k(\mathbf{z}_i)$ is the distance to the k-th nearest neighbor (*c.f.* Section 3).

Main result We show that our KNN-based OOD detector can reject inputs equivalent to the estimated Bayesian binary decision function. A small KNN distance $r_k(\mathbf{z}_i)$ directly translates into a high probability of being ID, and vice versa. We depict this in the following Theorem.

Theorem 6.1. *With the setup specified above, if* $\hat{p}_{out}(\mathbf{z}_i) = \hat{c}_0 \mathbf{1}\{\hat{p}_{in}(\mathbf{z}_i; k, n) < \frac{\beta\varepsilon\hat{c}_0}{(1-\beta)(1-\varepsilon)}\}$, *and* $\lambda = -\sqrt[m-1]{\frac{(1-\beta)(1-\varepsilon)k}{\beta\varepsilon c_b n \hat{c}_0}}$, *we have*

$$\mathbf{1}\{-r_k(\mathbf{z}_i) \geq \lambda\} = \mathbf{1}\{\hat{p}(g_i = 1|\mathbf{z}_i) \geq \beta\},$$

[3] In experiments, as it is difficult to simulate the universal OOD, we approximate it by using a diverse yet finite collection of datasets. Our theory is thus complementary to our experiments and captures the universality of OOD data.

[4] Note that β does not have to be $\frac{1}{2}$ for the Bayesian classifier to be optimal. β can be any value larger than $\frac{(1-\epsilon)c_1}{(1-\epsilon)c_1+\epsilon c_0}$ when $\epsilon c_0 \geq (1-\epsilon)c_1$.

where $\hat{p}(\cdot)$ denotes the empirical estimation. The proof is in Appendix A.

7. Related Work

OOD detection The phenomenon of neural networks' overconfidence in out-of-distribution data is first revealed in (Nguyen et al., 2015), which attracts growing research attention in several thriving directions:

(1) One line of work attempted to perform OOD detection by devising scoring functions, including OpenMax score (Bendale & Boult, 2015), maximum softmax probability (Hendrycks & Gimpel, 2017), ODIN score (Liang et al., 2018), deep ensembles (Lakshminarayanan et al., 2017), Mahalanobis distance-based score (Lee et al., 2018), energy score (Liu et al., 2020; Lin et al., 2021; Wang et al., 2021; Morteza & Li, 2022), activation rectification (ReAct) (Sun et al., 2021), gradient-based score (Huang et al., 2021) and ViM score (Wang et al., 2022). In Huang & Li (2021), the authors revealed that approaches developed for CIFAR datasets might not translate effectively into a large-scale ImageNet benchmark, and highlight the need to evaluate OOD detection methods in a real-world setting. To date, *none* of the prior works investigated the non-parametric nearest neighbor approach for OOD detection. Our work bridges the gap by presenting the first study exploring the efficacy of using nearest neighbor distance for OOD detection. We demonstrate superior performance on several OOD detection benchmarks, and we hope our work draws attention to the strong promise of the non-parametric approach.

(2) Another promising line of work addressed OOD detection by training-time regularization (Lee et al., 2017; Bevandić et al., 2018; Malinin & Gales, 2018; Hendrycks et al., 2019; Geifman & El-Yaniv, 2019; Hein et al., 2019; Meinke & Hein, 2019; Mohseni et al., 2020; Liu et al., 2020; Jeong & Kim, 2020; Van Amersfoort et al., 2020; Yang et al., 2021; Chen et al., 2021; Wei et al., 2022; Ming et al., 2022a; Katz-Samuels et al., 2022). For example, models are encouraged to give predictions with uniform distribution (Lee et al., 2017; Hendrycks et al., 2019) or higher energies (Liu et al., 2020; Ming et al., 2022a; Du et al., 2022a; Katz-Samuels et al., 2022) for outlier data. Most regularization methods require the availability of auxiliary OOD data. Recently, VOS (Du et al., 2022b) alleviates the need by automatically synthesizing virtual outliers that can meaningfully regularize the model's decision boundary during training.

(3) More recently, several works explored the role of representation learning for OOD detection. In particular, `CSI` (Tack et al., 2020) investigate the type of data augmentations that are particularly beneficial for OOD detection. Other works (Winkens et al., 2020; Sehwag et al., 2021) verify the effectiveness of applying the off-the-shelf multi-view contrastive losses such as SimCLR (Chen et al., 2020) and SupCon (Khosla et al., 2020) for OOD detection. These two works both use Mahalanobis distance as the OOD score, and make strong distributional assumptions by modeling the class-conditional feature space as multivariate Gaussian distribution. Ming et al. (2022b) propose a prototype-based contrastive learning framework for OOD detection, which promote stronger ID-OOD separability than SupCon loss. Our method and previous works are fundamentally different in the OOD detection method, despite all benefit from high-quality representations. In particular, KNN is a non-parametric method that does not impose prior of ID distribution. Performance-wise, our method outperforms SSD by a substantial margin, and is easy to use in practice.

KNN for anomaly detection KNN has been explored for anomaly detection (Jing et al., 2014; Zhao & Lai, 2020; Bergman et al., 2020), which aims to detect abnormal input samples from one class. We focus on OOD detection, which requires additionally performing multi-class classification for ID data. Some other recent works (Dang et al., 2015; Gu et al., 2019; Pires et al., 2020) explore the effectiveness of KNN-based anomaly detection for the tabular data. The potential of using KNN for OOD detection in deep neural networks is currently underexplored. Our work provides both new empirical insights and theoretical analysis of using the KNN-based approach for OOD detection.

8. Conclusion

This paper presents the first study exploring and demonstrating the efficacy of the non-parametric nearest-neighbor distance for OOD detection. Unlike prior works, the non-parametric approach does not impose any distributional assumption about the underlying feature space, hence providing stronger flexibility and generality. We provide important insights that a high-quality feature embedding and a suitable distance measure are two indispensable components for the OOD detection task. Extensive experiments show KNN-based method can notably improve the performance on several OOD detection benchmarks, establishing superior results. We hope our work inspires future research on using the non-parametric approach to OOD detection.

Acknowledgement

Work is supported by a research award from American Family Insurance. Zhu acknowledges NSF grants 1545481, 1704117, 1836978, 2041428, 2023239, ARO MURI W911NF2110317, and MADLab AF CoE FA9550-18-1-0166. The authors would also like to thank ICML reviewers for the helpful suggestions and feedback.

References

Bendale, A. and Boult, T. Towards open world recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1893–1902, 2015.

Bergman, L., Cohen, N., and Hoshen, Y. Deep nearest neighbor anomaly detection. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2020.

Bevandić, P., Krešo, I., Oršić, M., and Šegvić, S. Discriminative out-of-distribution detection for semantic segmentation. *arXiv preprint arXiv:1808.07703*, 2018.

Breunig, M. M., Kriegel, H.-P., Ng, R. T., and Sander, J. Lof: identifying density-based local outliers. In *Proceedings of the 2000 ACM SIGMOD International Conference on Management of Data*, pp. 93–104, 2000.

Chen, J., Li, Y., Wu, X., Liang, Y., and Jha, S. Atom: Robustifying out-of-distribution detection using outlier mining. *Proceedings of European Conference on Machine Learning and Principles and Practice of Knowledge Discovery in Databases*, 2021.

Chen, T., Kornblith, S., Norouzi, M., and Hinton, G. A simple framework for contrastive learning of visual representations. In *Proceedings of the International Conference on Machine Learning*, pp. 1597–1607. PMLR, 2020.

Cimpoi, M., Maji, S., Kokkinos, I., Mohamed, S., and Vedaldi, A. Describing textures in the wild. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 3606–3613, 2014.

Dang, T. T., Ngan, H. Y., and Liu, W. Distance-based k-nearest neighbors outlier detection method in large-scale traffic data. In *2015 IEEE International Conference on Digital Signal Processing (DSP)*, pp. 507–510. IEEE, 2015.

Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K., and Fei-Fei, L. Imagenet: A large-scale hierarchical image database. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 248–255, 2009.

Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J., and Houlsby, N. An image is worth 16x16 words: Transformers for image recognition at scale. In *Proceedings of the International Conference on Learning Representations*, 2021.

Du, X., Wang, X., Gozum, G., and Li, Y. 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*, 2022a.

Du, X., Wang, Z., Cai, M., and Li, S. Towards unknown-aware learning with virtual outlier synthesis. In *Proceedings of the International Conference on Learning Representations*, 2022b.

Geifman, Y. and El-Yaniv, R. Selectivenet: A deep neural network with an integrated reject option. *arXiv preprint arXiv:1901.09192*, 2019.

Gu, X., Akoglu, L., and Rinaldo, A. Statistical analysis of nearest neighbor methods for anomaly detection. In *Proceedings of the Advances in Neural Information Processing Systems*, volume 32, 2019.

He, K., Zhang, X., Ren, S., and Sun, J. Identity mappings in deep residual networks. In *Proceedings of the European Conference on Computer Vision*, pp. 630–645. Springer, 2016.

Hein, M., Andriushchenko, M., and Bitterwolf, J. 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*, pp. 41–50, 2019.

Hendrycks, D. and Gimpel, K. A baseline for detecting misclassified and out-of-distribution examples in neural networks. *Proceedings of the International Conference on Learning Representations*, 2017.

Hendrycks, D., Mazeika, M., and Dietterich, T. Deep anomaly detection with outlier exposure. *Proceedings of the International Conference on Learning Representations*, 2019.

Henze, N. and Zirkler, B. A class of invariant consistent tests for multivariate normality. *Communications in statistics-Theory and Methods*, 19(10):3595–3617, 1990.

Hsu, Y.-C., Shen, Y., Jin, H., and Kira, Z. Generalized odin: Detecting out-of-distribution image without learning from out-of-distribution data. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 10951–10960, 2020.

Huang, G., Liu, Z., Van Der Maaten, L., and Weinberger, K. Q. Densely connected convolutional networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 4700–4708, 2017.

Huang, R. and Li, Y. Mos: Towards scaling out-of-distribution detection for large semantic space. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 8710–8719, June 2021.

Huang, R., Geng, A., and Li, Y. On the importance of gradients for detecting distributional shifts in the wild. In *Proceedings of the Advances in Neural Information Processing Systems*, 2021.

Huber, P. J. Robust estimation of a location parameter. *Annals of Mathematical Statistics*, 35:73–101, March 1964.

Jeong, T. and Kim, H. Ood-maml: Meta-learning for few-shot out-of-distribution detection and classification. *Proceedings of the Advances in Neural Information Processing Systems*, 33:3907–3916, 2020.

Jing, T., Michael, A., and Pech, M. Anomaly detection using self-organizing maps-based k-nearest neighbor algorithm. In *PHM Society European Conference, 2(1)*, 2014.

Johnson, J., Douze, M., and Jégou, H. Billion-scale similarity search with gpus. *IEEE Transactions on Big Data*, 7 (3):535–547, 2019.

Katz-Samuels, J., Nakhleh, J., Nowak, R., and Li, Y. Training ood detectors in their natural habitats. In *International Conference on Machine Learning (ICML)*. PMLR, 2022.

Khosla, P., Teterwak, P., Wang, C., Sarna, A., Tian, Y., Isola, P., Maschinot, A., Liu, C., and Krishnan, D. Supervised contrastive learning. In *Proceedings of the Advances in Neural Information Processing Systems*, volume 33, pp. 18661–18673, 2020.

Lakshminarayanan, B., Pritzel, A., and Blundell, C. Simple and scalable predictive uncertainty estimation using deep ensembles. In *Proceedings of the Advances in Neural Information Processing Systems*, pp. 6402–6413, 2017.

Lee, K., Lee, H., Lee, K., and Shin, J. Training confidence-calibrated classifiers for detecting out-of-distribution samples. *arXiv preprint arXiv:1711.09325*, 2017.

Lee, K., Lee, K., Lee, H., and Shin, J. A simple unified framework for detecting out-of-distribution samples and adversarial attacks. In *Proceedings of the Advances in Neural Information Processing Systems*, pp. 7167–7177, 2018.

Liang, S., Li, Y., and Srikant, R. Enhancing the reliability of out-of-distribution image detection in neural networks. In *Proceedings of the International Conference on Learning Representations*, 2018.

Lin, Z., Roy, S. D., and Li, Y. Mood: Multi-level out-of-distribution detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 15313–15323, June 2021.

Liu, F. T., Ting, K. M., and Zhou, Z.-H. Isolation forest. In *2008 Eighth IEEE International Conference on Data Mining*, pp. 413–422, 2008. doi: 10.1109/ICDM.2008.17.

Liu, W., Wang, X., Owens, J., and Li, Y. Energy-based out-of-distribution detection. *Proceedings of the Advances in Neural Information Processing Systems*, 2020.

Loshchilov, I. and Hutter, F. Stochastic gradient descent with warm restarts. In *Proceedings of the International Conference on Learning Representations*, pp. 1–16, 2016.

Mahalanobis, P. C. On the generalized distance in statistics. National Institute of Science of India, 1936.

Malinin, A. and Gales, M. Predictive uncertainty estimation via prior networks. In *Proceedings of the Advances in Neural Information Processing Systems*, pp. 7047–7058, 2018.

McInnes, L., Healy, J., Saul, N., and Grossberger, L. Umap: Uniform manifold approximation and projection. *The Journal of Open Source Software*, 3(29):861, 2018.

Meinke, A. and Hein, M. Towards neural networks that provably know when they don't know. *Proceedings of the International Conference on Learning Representations*, 2019.

Ming, Y., Fan, Y., and Li, Y. Poem: Out-of-distribution detection with posterior sampling. In *Proceedings of the International Conference on Machine Learning*. PMLR, 2022a.

Ming, Y., Sun, Y., Dia, O., and Li, Y. Cider: Exploiting hyperspherical embeddings for out-of-distribution detection. *arXiv preprint arXiv:2203.04450*, 2022b.

Mohseni, S., Pitale, M., Yadawa, J., and Wang, Z. Self-supervised learning for generalizable out-of-distribution detection. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pp. 5216–5223, 2020.

Morteza, P. and Li, Y. Provable guarantees for understanding out-of-distribution detection. *Proceedings of the AAAI Conference on Artificial Intelligence*, 2022.

Netzer, Y., Wang, T., Coates, A., Bissacco, A., Wu, B., and Ng, A. Y. Reading digits in natural images with unsupervised feature learning. 2011.

Nguyen, A., Yosinski, J., and Clune, J. Deep neural networks are easily fooled: High confidence predictions for unrecognizable images. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 427–436, 2015.

Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., Killeen, T., Lin, Z., Gimelshein, N., Antiga, L., Desmaison, A., Kopf, A., Yang, E., DeVito, Z., Raison, M., Tejani, A., Chilamkurthy, S., Steiner, B., Fang, L., Bai, J., and Chintala, S. Pytorch: An imperative style, high-performance deep learning library. In *Proceedings of the Advances in Neural Information Processing Systems 32*, pp. 8024–8035. 2019.

Pevnỳ, T. Loda: Lightweight on-line detector of anomalies. *Machine Learning*, 102(2):275–304, 2016.

Pires, C., Barandas, M., Fernandes, L., Folgado, D., and Gamboa, H. Towards knowledge uncertainty estimation for open set recognition. *Machine Learning Knowledge*, 2:505–532, 2020.

Schölkopf, B., Platt, J. C., Shawe-Taylor, J., Smola, A. J., and Williamson, R. C. Estimating the Support of a High-Dimensional Distribution. *Neural Computation*, 13(7): 1443–1471, 07 2001. ISSN 0899-7667.

Sehwag, V., Chiang, M., and Mittal, P. Ssd: A unified framework for self-supervised outlier detection. In *Proceedings of the International Conference on Learning Representations*, 2021.

Shyu, M.-L., Chen, S.-C., Sarinnapakorn, K., and Chang, L. A novel anomaly detection scheme based on principal component classifier. Technical report, Miami Univ Coral Gables Fl Dept of Electrical and Computer Engineering, 2003.

Sun, Y., Guo, C., and Li, Y. React: Out-of-distribution detection with rectified activations. In *Proceedings of the Advances in Neural Information Processing Systems*, 2021.

Tack, J., Mo, S., Jeong, J., and Shin, J. Csi: Novelty detection via contrastive learning on distributionally shifted instances. In *Proceedings of the Advances in Neural Information Processing Systems*, 2020.

Van Amersfoort, J., Smith, L., Teh, Y. W., and Gal, Y. Uncertainty estimation using a single deep deterministic neural network. In *Proceedings of the International Conference on Machine Learning*, 2020.

Van Horn, G., Mac Aodha, O., Song, Y., Cui, Y., Sun, C., Shepard, A., Adam, H., Perona, P., and Belongie, S. The inaturalist species classification and detection dataset. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 8769–8778, 2018.

Wang, H., Liu, W., Bocchieri, A., and Li, Y. Can multi-label classification networks know what they don't know? *Proceedings of the Advances in Neural Information Processing Systems*, 2021.

Wang, H., Li, Z., Feng, L., and Zhang, W. Vim: Out-of-distribution with virtual-logit matching. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2022.

Wei, H., Xie, R., Cheng, H., Feng, L., An, B., and Li, Y. Mitigating neural network overconfidence with logit normalization. *Proceedings of the International Conference on Machine Learning*, 2022.

Winkens, J., Bunel, R., Roy, A. G., Stanforth, R., Natarajan, V., Ledsam, J. R., MacWilliams, P., Kohli, P., Karthikesalingam, A., Kohl, S., et al. Contrastive training for improved out-of-distribution detection. *arXiv preprint arXiv:2007.05566*, 2020.

Xiao, J., Hays, J., Ehinger, K. A., Oliva, A., and Torralba, A. Sun database: Large-scale scene recognition from abbey to zoo. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 3485–3492, 2010.

Xu, P., Ehinger, K. A., Zhang, Y., Finkelstein, A., Kulkarni, S. R., and Xiao, J. Turkergaze: Crowdsourcing saliency with webcam based eye tracking. *arXiv preprint arXiv:1504.06755*, 2015.

Yang, J., Wang, H., Feng, L., Yan, X., Zheng, H., Zhang, W., and Liu, Z. Semantically coherent out-of-distribution detection. In *Proceedings of the IEEE International Conference on Computer Vision*, pp. 8301–8309, October 2021.

Yu, F., Seff, A., Zhang, Y., Song, S., Funkhouser, T., and Xiao, J. Lsun: Construction of a large-scale image dataset using deep learning with humans in the loop. *arXiv preprint arXiv:1506.03365*, 2015.

Zhao, P. and Lai, L. Analysis of knn density estimation. *arXiv preprint arXiv:2010.00438*, 2020.

Zhou, B., Lapedriza, A., Khosla, A., Oliva, A., and Torralba, A. Places: A 10 million image database for scene recognition. *Proceedings of the IEEE transactions on pattern analysis and machine intelligence*, 40(6):1452–1464, 2017.

A. Theoretical Analysis

Proof of Theorem 6.1 We now provide the proof sketch for readers to understand the key idea, which revolves around performing the empirical estimation of the probability $\hat{p}(g_i = 1|\mathbf{z}_i)$. By the Bayesian rule, the probability of $\mathbf{z}$ being ID data is:

$$\begin{aligned} p(g_i = 1|\mathbf{z}_i) &= \frac{p(\mathbf{z}_i|g_i = 1) \cdot p(g_i = 1)}{p(\mathbf{z}_i)} \\ &= \frac{p_{in}(\mathbf{z}_i) \cdot p(g_i = 1)}{p_{in}(\mathbf{z}_i) \cdot p(g_i = 1) + p_{out}(\mathbf{z}_i) \cdot p(g_i = 0)} \\ \hat{p}(g_i = 1|\mathbf{z}_i) &= \frac{(1-\varepsilon)\hat{p}_{in}(\mathbf{z}_i)}{(1-\varepsilon)\hat{p}_{in}(\mathbf{z}_i) + \varepsilon\hat{p}_{out}(\mathbf{z}_i)}. \end{aligned}$$

Hence, estimating $\hat{p}(g_i = 1|\mathbf{z}_i)$ boils down to deriving the empirical estimation of $\hat{p}_{in}(\mathbf{z}_i)$ and $\hat{p}_{out}(\mathbf{z}_i)$, which we show below respectively.

Estimation for $\hat{p}_{in}(\mathbf{z}_i)$ Recall that $\mathbf{z}$ is a normalized feature vector in $\mathbb{R}^m$. Therefore $\mathbf{z}$ locates on the surface of a m-dimensional unit sphere. We denote $B(\mathbf{z}, r) = \{\mathbf{z}' : \|\mathbf{z}' - \mathbf{z}\|_2 \leq r\} \cap \{\|\mathbf{z}'\|_2 = 1\}$, which is a set of data points on the unit hyper-sphere and are at most r Euclidean distance away from the center $\mathbf{z}$. Note that the local dimension of $B(\mathbf{z}, r)$ is $m - 1$.

Assuming the density satisfies Lebesgue's differentiation theorem, the probability density function can be attained by:

$$p_{in}(\mathbf{z}_i) = \lim_{r \to 0} \frac{p(\mathbf{z} \in B(\mathbf{z}_i, r)|g_i = 1)}{|B(\mathbf{z}_i, r)|}.$$

In training time, we empirically observe n in-distribution samples $\mathbb{Z}_n = \{\mathbf{z}'_1, \mathbf{z}'_2, ..., \mathbf{z}'_n\}$. We assume each sample $\mathbf{z}'_j$ is *i.i.d* with a probability mass $\frac{1}{n}$. The empirical point-density for the ID data can be estimated by k-NN distance:

$$\begin{aligned} \hat{p}_{in}(\mathbf{z}_i; k, n) &= \frac{p(\mathbf{z}'_j \in B(\mathbf{z}_i, r_k(\mathbf{z}_i))|\mathbf{z}'_j \in \mathbb{Z}_n)}{|B(\mathbf{z}_i, r_k(\mathbf{z}_i))|} \\ &= \frac{k}{c_b n (r_k(\mathbf{z}_i))^{m-1}}, \end{aligned}$$

where c_b is a constant. The following Lemma A.1 establishes the convergence rate of the estimator.

Lemma A.1.

$$\lim_{\frac{k}{n} \to 0} \hat{p}_{in}(\mathbf{z}_i; k, n) = p_{in}(\mathbf{z}_i)$$

Specifically,

$$\mathbb{E}[|\hat{p}_{in}(\mathbf{z}_i; k, n) - p_{in}(\mathbf{z}_i)|] = o(\sqrt[m-1]{\frac{k}{n}} + \sqrt{\frac{1}{k}})$$

The proof is given in (Zhao & Lai, 2020).

Estimation for $\hat{p}_{out}(\mathbf{z}_i)$ A key challenge in OOD detection is the lack of knowledge on OOD distribution, which can arise universally outside ID data. We thus try to keep our analysis general and reflect the fact that we do not have any strong prior information about OOD. For this reason, we model OOD data with an equal chance to appear outside of the high-density region of ID data. Our theory is thus complementary to our experiments and captures the universality of OOD data. Specifically, we denote

$$\hat{p}_{out}(\mathbf{z}_i) = \hat{c}_0 \mathbf{1}\{\hat{p}_{in}(\mathbf{z}_i; k, n) < \frac{\beta\varepsilon\hat{c}_0}{(1-\beta)(1-\varepsilon)}\}$$

where the threshold is chosen to satisfy the theorem.

Lastly, our theorem holds by plugging in the empirical estimation of $\hat{p}_{in}(\mathbf{z}_i)$ and $\hat{p}_{out}(\mathbf{z}_i)$.

Proof.

$$
\begin{aligned}
\mathbf{1}\{-r_k(\mathbf{z}_i) \geq \lambda\} &= \mathbf{1}\{\varepsilon c_b n \hat{c}_0 (r_k(\mathbf{z}_i))^{m-1} \leq \frac{1-\beta}{\beta}(1-\varepsilon)k\} \\
&= \mathbf{1}\{\varepsilon c_b n \hat{c}_0 \mathbf{1}\{\varepsilon c_b n \hat{c}_0 (r_k(\mathbf{z}_i))^{m-1} > \frac{1-\beta}{\beta}(1-\varepsilon)k\}(r_k(\mathbf{z}_i))^{m-1} \leq \frac{1-\beta}{\beta}(1-\varepsilon)k\} \\
&= \mathbf{1}\{\varepsilon c_b n \hat{c}_0 \mathbf{1}\{\hat{p}_{in}(\mathbf{z}_i; k, n) < \frac{\beta\varepsilon\hat{c}_0}{(1-\beta)(1-\varepsilon)}\}(r_k(\mathbf{z}_i))^{m-1} \leq \frac{1-\beta}{\beta}(1-\varepsilon)k\} \\
&= \mathbf{1}\{\varepsilon c_b n \hat{p}_{out}(\mathbf{z}_i)(r_k(\mathbf{z}_i))^{m-1} \leq \frac{1-\beta}{\beta}(1-\varepsilon)k\} \\
&= \mathbf{1}\{\frac{k(1-\varepsilon)}{k(1-\varepsilon) + \varepsilon c_b n \hat{p}_{out}(\mathbf{z}_i)(r_k(\mathbf{z}_i))^{m-1}} \geq \beta\} \\
&= \mathbf{1}\{\hat{p}(g_i = 1|\mathbf{z}_i) \geq \beta\}
\end{aligned}
$$

□

B. Configurations

Non-parametric methods for anomaly detection We provide implementation details of the non-parametric methods in this section. Specifically,

`IForest` (Liu et al., 2008) generates a random forest assuming the test anomaly can be isolated in fewer steps. We use 100 base estimators in the ensemble and each estimator draws 256 samples randomly for training. The number of features to train each base estimator is set to 512.

`LOF` (Breunig et al., 2000) defines an outlier score based on the sample's k-NN distances. We set $k = 50$.

`LODA` (Pevnỳ, 2016) is an ensemble solution combining multiple weaker binary classifiers. The number of bins for the histogram is set to 10.

`PCA` (Shyu et al., 2003) detects anomaly samples with large values when mapping to the directions with small eigenvalues. We use 50 components for calculating the outlier scores.

`OCSVM` (Schölkopf et al., 2001) learns a decision boundary that corresponds to the desired density level set of with the kernel function. We use the RBF kernel with $\gamma = \frac{1}{512}$. The upper bound on the fraction of training error is set to 0.5.

Some of these methods (Schölkopf et al., 2001; Shyu et al., 2003) are specifically designed for anomaly detection scenarios that assume ID data is from one class. We show that k-NN distance with the class-aware embeddings can achieve both OOD detection and multi-class classification tasks.

C. Results on Different Architecture

In the main paper, we have shown that the nearest neighbor approach is competitive on ResNet. In this section, we show in Table 7 that KNN's strong performance holds on different network architectures DenseNet-101 (Huang et al., 2017). All the numbers reported are averaged over OOD test datasets described in Section 4.1.

Table 7. **Comparison results with DenseNet-101.** Comparison with competitive out-of-distribution detection methods. All methods are based on a model trained on ID data only. All values are percentages and are averaged over all OOD test datasets.

Method	**CIFAR-10**			**CIFAR-100**		
	FPR95 ↓	**AUROC** ↑	**ID ACC** ↑	**FPR95** ↓	**AUROC** ↑	**ID ACC** ↑
MSP	49.95	92.05	94.38	79.10	75.39	75.08
Energy	30.16	92.44	94.38	68.03	81.40	75.08
ODIN	30.02	93.86	94.38	55.96	85.16	75.08
Mahalanobis	35.88	87.56	94.38	74.57	66.03	75.08
GODIN	28.98	92.48	94.22	55.38	83.76	74.50
CSI	70.97	78.42	93.49	79.13	60.41	68.48
SSD+	16.21	96.96	**94.45**	43.44	88.97	**75.21**
KNN+ (ours)	**12.16**	**97.58**	**94.45**	**37.27**	**89.63**	**75.21**