How to Overcome Curse-of-Dimensionality for Out-of-Distribution Detection?

Soumya Suvra Ghosal[*], **Yiyou Sun***, **Yixuan Li**

Department of Computer Sciences, University of Wisconsin – Madison
{sghosal, sunyiyou, sharonli}@cs.wisc.edu

Abstract

Machine learning models deployed in the wild can be challenged by out-of-distribution (OOD) data from unknown classes. Recent advances in OOD detection rely on distance measures to distinguish samples that are relatively far away from the in-distribution (ID) data. Despite the promise, distance-based methods can suffer from the curse-of-dimensionality problem, which limits the efficacy in high-dimensional feature space. To combat this problem, we propose a novel framework, Subspace Nearest Neighbor ($\mathcal{S}$NN), for OOD detection. In training, our method regularizes the model and its feature representation by leveraging the most relevant subset of dimensions (i.e. subspace). Subspace learning yields highly distinguishable distance measures between ID and OOD data. We provide comprehensive experiments and ablations to validate the efficacy of $\mathcal{S}$NN. Compared to the current best distance-based method, $\mathcal{S}$NN reduces the average FPR95 by 15.96% on the CIFAR-100 benchmark.

1 Introduction

Modern machine learning models deployed in the wild face risks from out-of-distribution (OOD) data – samples from novel contexts and classes that were not taught to the model during training. Identifying OOD samples is paramount for safely operating machine learning models in uncertain environments. A safe system should be able to identify the unknown data as OOD so that the control can be passed to humans. This illuminates the importance of OOD detection, which allows the learner to express ignorance and take further precautions.

Recent solutions for OOD detection tasks often rely on distance measures to distinguish samples that are conspicuously dissimilar from the in-distribution (ID) data (Lee et al. 2018b; Tack et al. 2020; Sehwag, Chiang, and Mittal 2021; Sun et al. 2022). These methods leverage embeddings extracted from a well-trained classifier and operate under the assumption that test OOD samples are far away from the ID data in the feature space. For example, Sun et al. (2022) recently showed that non-parametric nearest neighbor distance is a strong indicator function for OOD likelihood. Despite the promise, distance measures can be sensitive to the dimensionality of the feature space. Those research questions have been raised in the

90s, including Beyer et al. (1999), as to whether the nearest neighbor is meaningful in high dimensions. The key result of (Beyer et al. 1999) states that in high dimensional spaces, the difference between the minimum and the maximum distance between a random reference point and a list of n random data points $\mathbf{x}_1,...,\mathbf{x}_n$ become indiscernible compared to the minimum distance:

$$\lim_{d \to \infty} E\left(\frac{D_{\max} - D_{\min}}{D_{\min}}\right) \to 0. \tag{1}$$

This phenomenon—recognized as one aspect of the "*curse of dimensionality*" (Beyer et al. 1999; Hinneburg, Aggarwal, and Keim 2000; Aggarwal, Hinneburg, and Keim 2001; Houle et al. 2010; Kriegel et al. 2009)—limits the efficacy of distance-based OOD detection method for modern neural networks. For example, let's consider 50000 points sampled randomly from a 2048 dimensional space bounded by $[0, 1]$. Given a test sample, to find the 10-th nearest neighbor, we need a hyper-cube with a length of at least 0.9958[1], which covers the entire feature space. In other words, the high-dimensional spaces can lead to instability of nearest neighbor queries. Despite its importance, the problem has received little attention in the literature on OOD detection. This begs the following question:

How do we combat the curse-of-dimensionality for OOD detection?

Targeting this important problem, we propose a new framework called *Subspace Nearest Neighbor* (**SNN**) for OOD detection. Our key idea is to learn a feature subspace—a subset of relevant features—for distance-based detection. Our method is motivated by the observation drawn in Houle et al. (2010): irrelevant attributes in a feature space may impede the separability of different distributions and thus have the potential for rendering the nearest neighbor less meaningful. In light of this, $\mathcal{S}$NN regularizes the model and its feature representation by leveraging the most relevant subset of dimensions (*i.e.* subspace) for the class prediction. Geometrically, this is equivalent to projecting the embedding to the selected subspace, and the output for each class is derived using the projected features. The subspaces are learned in a way that maintains the discriminative power for classifying ID data.

---

[*]Equal contributions

[1]The calculation is done by $0.9958^{2048} \approx \frac{10}{50000}$.

During testing, distance-based OOD detection is performed by leveraging the learned features.

We show that $\mathcal{S}$NN is both theoretically grounded (Section 4) and empirically effective (Section 5) to combat the curse-of-dimensionality for OOD detection. Extensive experiments show that $\mathcal{S}$NN substantially outperforms the competitive methods in the literature. For example, using the CIFAR-100 dataset as ID and LSUN (Yu et al. 2015) as OOD data, our approach reduces the FPR95 from 66.09% to 24.43%—a **41.66**% direct improvement over the current best distance-based method KNN (Sun et al. 2022). Beyond the OOD detection task, our subspace learning algorithm also leads to improved calibration performance of the ID data itself[2]. We summarize our contributions below:

1. We propose a new framework, *subspace nearest neighbor* ($\mathcal{S}$NN), for distance-based OOD detection. $\mathcal{S}$NN effectively alleviates the curse-of-dimensionality suffered by the vanilla KNN approach operating on the full feature space.

2. We demonstrate the strong performance of $\mathcal{S}$NN on an extensive collection of OOD detection tasks. On CIFAR-100, $\mathcal{S}$NN substantially reduces the average FPR95 by **15.96**% compared to the current best distance-based approach (Sun et al. 2022). Further, we provide in-depth analysis and ablations to understand the effect of each component in our method (Section 6).

3. We provide theoretical analysis, showing that ID and OOD data become more distinguishable in a subspace with reduced dimensions. We hope our work draws attention to the strong promise of subspace learning for OOD detection.

2 Preliminaries

Setup. In this paper, we consider a supervised multi-class classification problem, where $\mathcal{X}$ denotes the input space and $\mathcal{Y} = \{1, 2, ..., C\}$ denotes the label space. The training set $\mathbb{D}_{\text{in}} = \{(\mathbf{x}_i, y_i)\}_{i=1}^{N}$ 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. Our framework concerns a common real-world scenario in which the algorithm is trained on the ID data but will then be deployed in environments containing *out-of-distribution* samples from unknown class $y \notin \mathcal{Y}$ and therefore should not be predicted by f. Formally, OOD detection can be formulated as a level-set estimation problem. At test time, one can perform the following test to determine whether a sample $\mathbf{x} \in \mathcal{X}$ is from $\mathcal{P}_{\text{in}}$ (ID) or not (OOD):

$$G_\lambda(\mathbf{x}) = \begin{cases} \text{ID} & \text{if } S(\mathbf{x}) \geq \lambda, \\ \text{OOD} & \text{if } S(\mathbf{x}) < \lambda \end{cases} \quad (2)$$

where $S(\mathbf{x})$ denotes a scoring function and λ is a chosen threshold such that a high fraction (95%) of ID data is correctly classified. Test samples with higher values of S($\mathbf{x}$) are classified as ID and vice-versa.

---

[2]Code is available at https://github.com/deeplearning-wisc/SNN

3 Methodology

A common design of $S(\mathbf{x})$ often relies on distance measures (Sehwag, Chiang, and Mittal 2021; Sun et al. 2022; Lee et al. 2018b) to distinguish OOD samples that are far away from the ID data in the feature space. Despite the promise, they operate on the full feature space with a large dimension, which is known to be susceptible to the "curse-of-the dimensionality". Intuitively, irrelevant attributes in a feature space may impede the separability of different distributions and thus have the potential to render the distance measure less meaningful. Rather than relying on the full feature space, our key idea is to use a subset of relevant dimensions for distance-based OOD detection.

In what follows, we formally introduce our framework, *Subspace Nearest Neighbor* (dubbed $\mathcal{S}$NN) for OOD detection. $\mathcal{S}$NN consists of two components. First, during training, the model learns the relevant subset of dimensions (*i.e.* subspace) for each class. The subspaces are learned in a way that maintains the discriminative power for classifying ID data (Section 3.1). Second, during testing, distance-based OOD detection is performed by using the learned features (Section 3.2).

3.1 Learning the Subspaces

We start with the first challenge: how to learn the relevant subset of dimensions (a.k.a. subspace)? Our idea is driven by the fact that the class prediction of a given example depends only on a subset of relevant features. Moreover, the relevance of any particular feature dimension may vary across different classes and instances. For example, for a DOG class instance, features such as tail, nose, etc are relevant. However, these feature dimensions can be irrelevant for an instance from BIRD class. This necessitates training the model using *class-dependent subspaces* conditioned on the given instance. We formally describe the learning procedure in the sequel.

Defining the subspace. We consider a deep neural network parameterized by θ, which encodes an input $\mathbf{x} \in \mathcal{X}$ to a feature space with dimension m. We denote by $h(\mathbf{x}) = [h^{(1)}(\mathbf{x}), h^{(2)}(\mathbf{x}), ..., h^{(m)}(\mathbf{x})]$ the feature vector from the penultimate layer of the network, which is a m-dimensional vector. For a class c, we select the subspace by masking the original feature:

$$h_{\text{masked}}(\mathbf{x}) = h(\mathbf{x}) \odot R_c(\mathbf{x}), s.t., R_c(\mathbf{x}) \in \{0, 1\}^m, \|R_c(\mathbf{x})\|_1 = s, \quad (3)$$

where $\odot$ represents the element-wise multiplication and $R_c(\mathbf{x})$ is a binary mask with s non-zero elements encoding the subset of dimensions (*i.e.* subspace) for class c. This way, the high-dimensional features can be projected onto the corresponding subspace defined by $R_c(\mathbf{x})$, which is class-dependent for a given input $\mathbf{x}$. The model output $f_c(\mathbf{x}; \theta, R_c(\mathbf{x}))$ is further obtained by passing the projected feature through a linear transformation with weight vector $\mathbf{w}_c \in \mathbb{R}^m$:

$$f_c(\mathbf{x}; \theta, R_c(\mathbf{x})) = \langle \mathbf{w}_c, h(\mathbf{x}) \odot R_c(\mathbf{x}) \rangle. \quad (4)$$

Learning subspace via relevance. Given an input $\mathbf{x}$, we now describe how to identify $R_c(\mathbf{x})$, the subset of most *relevant* dimensions for a class c. Our key idea is to formulate

the following optimization objective:

$$\min_{\theta} \ \mathbb{E}_{(\mathbf{x},y)\sim\mathcal{P}_{\mathcal{X}\mathcal{Y}}} \left[\mathcal{L}_{\text{CE}} \left(f(\mathbf{x};\theta, R_c(\mathbf{x})), y \right) \right], \qquad (5)$$

where $\quad f_c(\mathbf{x};\theta, R_c(\mathbf{x}))$
$$= \max_{R_c(\mathbf{x})\in\{0,1\}^m, \|R_c(\mathbf{x})\|_1=s} \langle \mathbf{w}_c, h(\mathbf{x}) \odot R_c(\mathbf{x}) \rangle. \quad (6)$$

Equation 6 indicates that the subspace is chosen based on the features that are most responsible for the class prediction. Geometrically, this is equivalent to projecting the point to the selected subspace, and the output for each class is derived using the projected features. The search of subspace can be computed efficiently. One can easily prove that $f_c(\mathbf{x};\theta, R_c(\mathbf{x}))$ in Equation 6 is equivalent to the summation of the top s entries in $\mathbf{w}_c \odot h(\mathbf{x})$ with the highest values. In Section 5.1, we show that our subspace learning objective incurs similar (and in some cases faster) training time compared to the standard cross-entropy loss. During inference, given a test sample $\mathbf{x}'$, the class prediction $\hat{y}$ is made by $\hat{y} = \arg\max_{c\in\mathcal{Y}} \langle \mathbf{w}_c, h(\mathbf{x}') \odot R_c(\mathbf{x}') \rangle$.

Our formulation also flexibly allows each feature dimension to be used for multiple class predictions. As an example, a class LAPTOP might depend on the two most relevant feature dimensions: KEYBOARD and SCREEN. Similarly, the class prediction for TV might be relying on two features TELECONTROL and SCREEN. In both cases, the feature of SCREEN is responsible for both TV and Laptop.

3.2 Out-of-Distribution Detection with Subspace Nearest Neighbor

We now describe how to perform test-time OOD detection leveraging the learned features. In particular, given a test sample's embedding $\mathbf{z}'$, we determine its OOD-ness by computing the k-th nearest neighbor distance *w.r.t.* the training embeddings. Here $\mathbf{z}' = h(\mathbf{x}')/\|h(\mathbf{x}')\|_2$ is the L_2-normalized feature embedding. The decision function for OOD detection is given by:

$$G(\mathbf{x}';k,\mathbb{D}_{\text{in}}) = \mathbf{1}\{-r_k(\mathbf{z}') \geq \lambda\},$$

where $r_k(\mathbf{z}') = ||\mathbf{z}' - \mathbf{z}_k||_2$ is the distance to the k-th nearest neighbor (k-NN) training embedding ($\mathbf{z}_k$) 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. Different from Sun et al. (2022), we calculate the nearest neighbor distance based on the subspace-regularized feature space, rather than the vanilla ERM-trained model. Next, we show that our subspace nearest neighbor approach is both theoretically grounded (Section 4) and empirically effective (Section 5) to combat the curse-of-dimensionality for OOD detection.

4 Theoretical Insights

In this section, we provide theoretical insights into why using a feature subspace is critical for k-NN distance-based OOD detection.

Setup. We consider the OOD detection task as a special binary classification task, where the negative samples (OOD)

are only available in the testing stage. A key challenge in OOD detection (and theoretical analysis) is the lack of knowledge on OOD distribution. Common algorithms detect OOD samples when ID density $p(\mathbf{z}_i|\mathbf{z}_i \in \text{ID})$ is low. For simplicity, we let $p_{in}(\mathbf{z}_i) = p(\mathbf{z}_i|\mathbf{z}_i \in \text{ID})$. We thus divide the bounded-set $\mathcal{Z}$ into two disjoint set $\mathcal{Z}_{in}$ and $\mathcal{Z}_{out}$ with:

$$\mathcal{Z}_{in} = \{\mathbf{z}|p_{in}(\mathbf{z}) \geq \lambda\}, \mathcal{Z}_{out} = \{\mathbf{z}|p_{in}(\mathbf{z}) < \lambda\}, \quad (7)$$

where λ is a certain real number.

Main result. In our analysis, we aim to investigate the impact of embedding dimension m on the following gap:

$$\hat{\Delta}(m) = \mathbb{E}[\hat{p}_{in}(\mathbf{z})|\mathbf{z} \in \mathcal{Z}_{in}] - \mathbb{E}[\hat{p}_{in}(\mathbf{z})|\mathbf{z} \in \mathcal{Z}_{out}], \quad (8)$$

which is the difference of average estimated density between $\mathcal{Z}_{in}$ and $\mathcal{Z}_{out}$. This gap can be translated into OOD detection performance with the k-NN distance. To see this, a function of the k-NN distance $r_k(\mathbf{z})$ can be regarded as an estimator which approximates $p_{in}(\mathbf{z})$ (Zhao and Lai 2022):

$$\hat{p}_{in}(\mathbf{z}) = \frac{k-1}{NV(B(\mathbf{z}, r_k(\mathbf{z}))}, \quad (9)$$

where $V(B(\mathbf{z}, r_k(\mathbf{z})))$ denotes the volume of the ball $B(\mathbf{z}, r_k(\mathbf{z}))$ centered at $\mathbf{z}$ with radius $r_k(\mathbf{z})$ and N represents number of training samples. To investigate how OOD detection performance with k-NN distance is impacted by the curse-of-dimensionality, we derive the following theorem showing how the lower bound of $\hat{\Delta}(m)$ changes when m becomes large.

Theorem 4.1. *Let* $\mathbb{E}[p_{in}(\mathbf{z})|\mathbf{z} \in \mathcal{Z}_{in}] - \mathbb{E}[p_{in}(\mathbf{z})|\mathbf{z} \in \mathcal{Z}_{out}] = \Delta(m)$ *be a function of the feature dimensionality* m. *We have the following bound (proof is in Appendix B):*

$$\hat{\Delta}(m) \geq \Delta(m) - O\left(\left(\frac{k}{N}\right)^{\frac{1}{m}} + k^{-\frac{1}{2}} \right). \quad (10)$$

Implications of theory. In Theorem 4.1, we see that $\hat{\Delta}(m)$ is lower-bounded by two terms: $\Delta(m)$ and $-O\left(\left(\frac{k}{N}\right)^{\frac{1}{m}} + k^{-\frac{1}{2}} \right)$. Specifically, $\Delta(m)$ represents the true density gap between the ID set and OOD set, which generally sets a ground-truth oracle on how well we can perform OOD detection. The second term $-O\left(\left(\frac{k}{N}\right)^{\frac{1}{m}} + k^{-\frac{1}{2}} \right)$ measures the approximation error when using k-NN distance to estimate the true density gap. The approximation error worsens when m gets larger, which is exactly where the "curse of dimensionality" comes from. Consider the following case: when $\Delta(m)$ is nearly constant, $-O\left(\left(\frac{k}{N}\right)^{\frac{1}{m}} + k^{-\frac{1}{2}} \right)$ monotonically decreases when m gets larger, then it is likely for $\hat{\Delta}(m)$ goes to 0 (indistinguishable between ID & OOD). This theoretical finding further validates the necessity of using a subspace of embedding, which is exactly the solution provided by $\mathcal{S}$NN.

Remark on a small m. If we only focus on the second term $-O\left(\left(\frac{k}{N}\right)^{\frac{1}{m}} + k^{-\frac{1}{2}} \right)$, it seems like we need the dimension m to be as small as possible. However, in practice, a very

Methods	Far-OOD		Near-OOD	
	FPR95 ↓	AUROC ↑	FPR95 ↓	AUROC ↑
MSP	52.11	91.79	64.66	85.28
ODIN	26.47	94.48	52.32	88.90
GODIN	17.42	95.84	60.69	82.37
Energy score	28.40	94.22	50.64	88.66
ReAct	33.12	94.32	53.51	88.96
GradNorm	24.79	92.58	65.44	79.31
LogitNorm	19.61	95.51	55.08	88.03
DICE	20.83	95.24	58.60	87.11
Mahalanobis	44.55	82.56	87.71	78.93
KNN	18.50	96.36	58.34	87.90
$\mathcal{S}$NN (ours)	**14.99** ±0.87	**97.15** ±0.27	**50.10** ±1.09	**89.80** ±0.65

Table 1: Performance comparison on near-OOD and far-OOD detection task. Architecture used is DenseNet-101 and ID data is CIFAR-10. We report the mean and variance across 3 training runs.

Method	FPR95 ↓	AUROC ↑	ID Acc. ↑
MSP	77.59	76.47	75.14
ODIN	56.39	86.02	75.14
GODIN	44.08	89.05	74.37
Energy score	57.07	84.83	75.14
ReAct	75.06	79.51	66.56
GradNorm	63.05	79.80	75.14
LogitNorm	61.10	84.72	75.42
DICE	49.72	87.23	68.65
Mahalanobis	56.93	80.27	75.14
KNN	47.21	85.27	75.14
$\mathcal{S}$NN (ours)	**31.25** ±1.25	**90.76** ±0.36	**75.59** ±0.08

Table 2: Performance comparison on CIFAR-100 dataset. We use DenseNet-101 for all baselines. We report the mean and variance across 3 different training runs.

small dimension will lose the necessary information needed to separate ID and OOD. Such implications are captured by $\Delta(m)$ which measures the true density gap between ID and OOD. If $\Delta(m)$ is an increasing function over m, there is a trade-off in choosing m with the second term that inversely encourages a smaller m. Such insight is indeed verified empirically in Section 6 that there exists an optimal m— neither too small nor too large.

5 Experiments

In this section, we extensively evaluate the effectiveness of our proposed method. The goal of our experimental sections is to mainly answer the following questions: (1) Can $\mathcal{S}$NN alleviate the curse of dimensionality? (2) How does $\mathcal{S}$NN compare against the state-of-the-art OOD detection methods? Due to space constraints, extensive experimental details are in Appendix C. Our code is open-sourced for the research community.

5.1 Evaluation on Common Benchmarks

Datasets. In this section, we make use of commonly studied CIFAR-10 (10 classes) and CIFAR-100 (100 classes) (Krizhevsky, Hinton et al. 2009) datasets as ID. Both datasets consist of images of size 32×32. We use the standard split with $50,000$ images for training and $10,000$ images for testing. We evaluate the methods on common OOD datasets: Textures (Cimpoi et al. 2014), SVHN (Netzer et al. 2011), LSUN-Crop (Yu et al. 2015), LSUN-Resize (Yu et al. 2015), iSUN (Xu et al. 2015), and Places365 (Zhou et al. 2017). Images in all these test datasets are of size 32×32.

Evaluation metrics. We compare the performance of various methods using the following metrics: (1) FPR95 measures the false positive rate (FPR) of OOD samples when 95% of ID samples are correctly classified; (2) AUROC is

the area under the Receiver Operating Characteristic curve; and (3) ID Acc. measures the ID classification accuracy.

Comparison with competitive methods. In Table 2, we provide a comprehensive comparison with competitive OOD detection baselines on CIFAR-100. We provide a detailed description of baseline approaches in Appendix C.3. We observe that our proposed method $\mathcal{S}$NN significantly outperforms the latest rivals. For a fair comparison, we divide the baselines into two categories: methods using model outputs and methods using feature representations. From Table 2, we highlight two salient observations: (1) Considering methods based on feature representations, $\mathcal{S}$NN outperforms KNN (non-parametric) and Mahalanobis (parametric) by **15.96%** and **25.68%** respectively in terms on FPR95. The results validate that learning feature subspace effectively alleviates the "curse-of-dimensionality" problem that is troubling the existing KNN approach. (2) Further, $\mathcal{S}$NN also performs better than output-based methods such as ReAct (Sun, Guo, and Li 2021). Specifically, with CIFAR-100 as ID, $\mathcal{S}$NN provides a **43.81%** improvement in FPR95 as compared to ReAct (Sun, Guo, and Li 2021). Notably, $\mathcal{S}$NN provides a **18.47%** improvement compared to (Sun and Li 2022), a post-hoc sparsification method. While DICE can severely affect the ID test accuracy (68.65%), $\mathcal{S}$NN exhibits stronger classification performance (75.59%) by baking in the inductive bias of subspaces through training. An extensive discussion is provided in Section 7.

Evaluation on near-OOD data. In Table 1, we compare the performance in detecting near-OOD data, which refers to samples near the ID data. Near-OOD is particularly challenging to detect, and can often be misclassified as ID. We report the performance on CIFAR-10 (ID) vs. CIFAR-100 (OOD), which is the most commonly used dataset pair for this task. We observe that $\mathcal{S}$NN consistently outperforms existing algorithms for near-OOD detection tasks, further demonstrating its strengths. Compared to KNN, $\mathcal{S}$NN reduces the FPR95 by 8.24%. For completeness, we also provide far-OOD evalu-

Method	Dataset (ID)	FPR95 ↓	AUROC ↑
Mahalanobis	CIFAR-10	44.55	82.56
$\mathcal{S}$NN (w. Mahalanobis)	CIFAR-10	**34.68**	**87.87**
Mahalanobis	CIFAR-100	56.93	80.27
$\mathcal{S}$NN (w. Mahalanobis)	CIFAR-100	**55.05**	**80.77**

Table 3: $\mathcal{S}$NN is also compatible with parametric approaches such as Mahalanobis distance (Lee et al. 2018b). The model is DenseNet. All values are averaged over six OOD test datasets.

Training Method	CIFAR-10	CIFAR-100
	(Train time in hours)	
Standard	2.10	2.25
$\mathcal{S}$NN	1.75	1.89

Table 4: Computational cost for training. Model used is DenseNet-101. For the comparison, we used the software configuration as reported in Appendix C.2.

ation results on CIFAR-10, where $\mathcal{S}$NN achieves an average FPR95 of 14.99%. Full result on each test dataset for CIFAR-10 is available in Appendix D.4.

Compatibility with other distance-based approaches. Beyond KNN (Sun et al. 2022), the Mahalanobis distances (Lee et al. 2018b) is also one of the most popular distance-based approaches to detect OOD. However, all prior solutions measure the distance with a full feature space which can also suffer from the curse of dimensionality. In this section, we show that subspace learning can also benefit parametric approaches like Mahalanobis distance (Lee et al. 2018b). In Table 3, we compare the OOD detection performance of using Mahalanobis distance on the vanilla model and the model trained with $\mathcal{S}$NN. We see that coupling subspace learning (in training) with Mahalanobis distance (in testing) reduces FPR95 by 9.87% and 1.88% on CIFAR-10 and CIFAR-100 datasets respectively.

Computational complexity. In Table 4, we compare the training time of $\mathcal{S}$NN with the standard training method using cross-entropy loss. We observe that training using $\mathcal{S}$NN incurs no additional computation overhead but rather is slightly more efficient compared to standard training procedures. This is because we perform gradient descent only on a subset of weights corresponding to the selected feature subspace. Thus, our method overall leads to faster updates and convergence. In Appendix D.1, we further show that $\mathcal{S}$NN remains competitive and outperforms the KNN counterpart on other common architecture.

6 Further Understanding of $\mathcal{S}$NN

Through comprehensive evaluations in Section 5, we have established the effectiveness of $\mathcal{S}$NN for OOD detection. In this section, we provide an in-depth analysis of several questions: (1) How does the subspace dimension impact the performance? (2) How does OOD detection performance

change if we change the subspace selection strategy? (3) How does the k in k-NN distance impact the OOD detection performance? We show comprehensive ablation studies to answer these questions. For consistency, all ablations are based on CIFAR-100 as ID dataset and DenseNet (Huang et al. 2017) architecture unless specified otherwise.

Ablation on subspace dimension. In this ablation, we aim to empirically verify our theoretical analysis in Section 4, and understand the effect of subspace dimension. We start by defining the **relevance ratio** $r = \frac{s}{m} \in (0, 1]$, which captures the sparsity of feature space used in training $\mathcal{S}$NN. Recall that m is the original dimension of features and s is the dimension we kept in $\mathcal{S}$NN.

In simple words, r represents the ratio between the dimension of the subspace and the full feature dimension. Specifically, we train and compare multiple models by varying $r = \{0.05, 0.15, 0.25, 0.35, 0.55, 0.75\}$. Figure 1 (left) summarizes the effect of r on OOD detection. We observe that: (1) Irrespective of the relevance ratio used, $\mathcal{S}$NN is consistently better than KNN (Sun et al. 2022) when $r < 1$. This validates the efficacy of learning feature subspaces for OOD detection, without excessive hyperparameter tuning. (2) Setting a mild ratio (e.g. $r = 0.25$) provides the optimal OOD detection performance, which is consistent with one chosen using our validation strategy (see Appendix C.4). (3) In the extreme case, when r is too small (e.g. $r = 0.05$), we observe a deterioration in the OOD detection performance. Overall our empirical observations align well with our theoretical insight provided in Section 4.

Visualization of the learned weight matrix. To further verify our method, we visualize in Figure 2 the learned final-layer weight matrix under different relevance ratios ($r = s/m$). For each r, we visualize the 342 dimensional weight vector corresponding to each class in CIFAR-100, i.e., a 342×100 matrix. When $r = 1$ (i.e. without any subspace constraint), the model utilizes the full feature space. Further, the visualization confirms that decreasing the relevance ratio effectively reduces feature subspace dimensionality.

Method	FPR95 ↓	AUROC ↑
Random subspace (Ho 1998)	42.21	84.97
Subspace with least relevance	63.96	80.82
$\mathcal{S}$NN (ours)	**31.25**	**90.76**

Table 5: Ablation on subspace selection methods. Model is DenseNet. All values are averaged over multiple OOD test datasets.

Ablation on subspace selection. A core component of our algorithm involves selecting the *most relevant* dimension for a class prediction. In particular, the subspace is chosen based on the dimensions that contributed most to the model's output. In this ablation, we contrast our subspace selection mechanism with random subspace (Ho 1998), a classical alternative. The random subspace relies on a stochastic process that randomly selects s components in the feature vector.
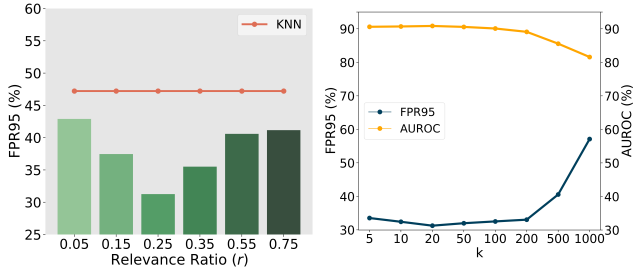
Figure 1: Left: Effect of varying the relevance ratio (r) on OOD detection performance when $k = 20$. Right: Effect of varying the number of neighbors (k) when $r = 0.25$.
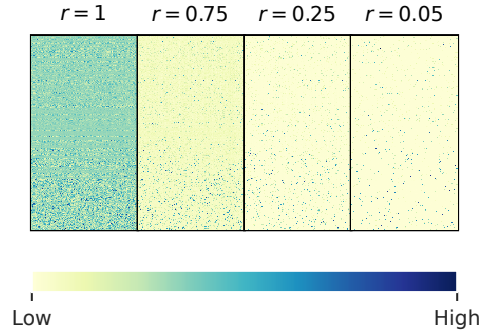

Figure 2: Visualization of learned final-layer weight matrix for $r \in \{1, 0.75, 0.25, 0.05\}$. For each r, we visualize a 342-dimensional weight vector corresponding to each class in CIFAR-100.

Simply put, this approach randomly sets s out of m elements in each R_c vector to be 1 and 0 elsewhere. We report ablation results in Table 5. For a fair comparison, we use relevance ratio $r = 0.25$ and the number of neighbors $k = 20$ for all methods. Empirical results highlight that randomly chosen subsets of dimensions are sub-optimal for the OOD detection task. Lastly, we also contrast with selecting the *least relevant* feature dimensions. We replace Equation 6 with:

$$f_c(\mathbf{x}; \theta, R_c(\mathbf{x}))$$
$$= \min_{R_c(\mathbf{x}) \in \{0,1\}^m, \|R_c(\mathbf{x})\|_1 = s} \langle \mathbf{w}_c, h(\mathbf{x}) \odot R_c(\mathbf{x}) \rangle, \quad (11)$$

which essentially changes from MAX to MIN. As expected, using the least relevant feature dimension results in significantly worse OOD detection performance.

Ablation on number of nearest neighbors k. In Figure 1 (right), we visualize the effect of varying the number of nearest-neighbors (k) on OOD detection performance. Here the model is trained with $r = 0.25$ on CIFAR-100. Specifically, we vary $k \in \{5, 10, 20, 50, 100, 200, 500, 1000\}$. We observe that the OOD detection performance is relatively stable under a mild k. In Appendix D.2, we further visualize the interaction between the two hyper-parameters r and k through OOD detection performance.

$\mathcal{S}$**NN improves calibration performance.** In addition to superior OOD detection performance, we aim to further investigate the calibration performance of ID data itself. As a quick recap, the calibration performance measures the alignment between the model's confidence and its actual predictive accuracy. We hypothesize that learning the feature subspace helps alleviate the problem of over-confident predictions on ID data, thereby improving model calibration. We verify in Appendix D.3 that training with $\mathcal{S}$NN indeed significantly improves the model calibration.

$\mathcal{S}$**NN scales to large datasets.** In this section, we evaluate $\mathcal{S}$NN on a more realistic high-resolution dataset ImageNet (Deng et al. 2009). Compared to CIFAR-100, inputs scale up in size in ImageNet-100 (we follow standard data augmentation pipelines and resize the input to 224 by 224). For
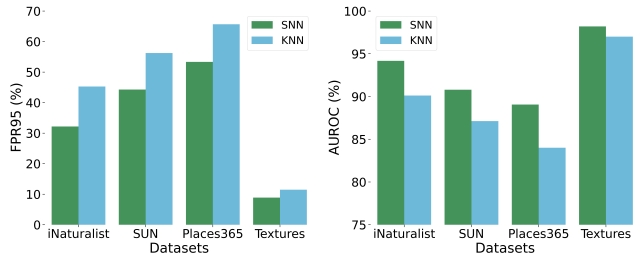

Figure 3: OOD detection performance comparison on ImageNet dataset (ID). $\mathcal{S}$NN consistently outperforms KNN across all OOD test datasets on the same architecture.

OOD test datasets, we use the same ones in (Huang and Li 2021), including subsets of `Places365`, `Textures`, `iNaturalist` and `SUN`. We train the ResNet-101 model for 100 epochs using a batch size of 256, starting from randomly initialized weights. We use SGD with a momentum of 0.9, and a weight decay of 1e-4. We set the initial learning rate as 0.1 and use a cosine-decay schedule. We set $r = 0.35$ and $k = 200$ based on our validation strategy described in Appendix C.4. We contrast two models trained with and without subspace learning. The results in terms of FPR95 and AUROC are shown in Figure 3. The results suggest that $\mathcal{S}$NN remains effective on all the OOD test sets and consistently outperforms KNN. This further verifies the benefits of explicitly promoting feature subspace to combat curse-of-dimensionality.

7 Discussion

Relations to DICE (Sun and Li 2022). Our work differs from DICE in two crucial aspects, both in terms of training loss and test-time OOD detection mechanism. As introduced in Section 3, $\mathcal{S}$NN can be viewed as *training-time* regularization with subspace. In contrast, DICE (Sun and Li 2022) only explored *test-time* weight sparsification mechanism, without explicitly learning the subspace in training. Specifically, DICE employs the standard cross-entropy loss, whereas we

contribute a new learning objective (*c.f.* Section 3.1) for OOD detection. Importantly, we show that training-time regularization provides substantial gains over simple post-hoc sparsification. For example, as evidenced in Table 2, post-hoc sparsification can severely affect the ID test accuracy. In contrast, our method bakes the inductive bias of "feature subspace" into training time, and exhibits stronger ID generalization in testing time. Besides, another major difference between $\mathcal{S}$NN and DICE lies in how test-time OOD detection is performed. In particular, $\mathcal{S}$NN is a distance-based method that operates in the *feature* space, whilst DICE derives OOD scores from the model *output* space. Different from DICE, $\mathcal{S}$NN is motivated to address the curse-of-dimensionality issue for OOD detection. Table 2 compares the OOD detection performance of $\mathcal{S}$NN with DICE. For example, on CIFAR-100, $\mathcal{S}$NN reduces FPR95 by **18.47%** compared to DICE. Improved performance validates the advantage of learning a subspace in training as opposed to test-time sparsification.

Relations to Dropout strategies. A commonly used technique to prevent over-parameterization is Unit Dropout (Hinton et al. 2012; Srivastava et al. 2014). Unlike Random Subspace (Ho 1998), Unit Dropout randomly sets elements in the feature activation vector $h(\mathbf{x})$ to be 0. Ba and Frey (2013) propose Adaptive Dropout, which generalizes the prior approach by learning the dropout probability using a binary belief network. Gomez et al. (2019) proposed Targeted Dropout which is based on keeping top-k weights in a fully connected layer with the highest magnitude. Their motivation is that low-valued weights should be allowed to increase their value during training, if they become important. Hence, instead of completely pruning the rest of the weights, they apply dropout with a fixed rate to randomly drop a fraction of the low-valued weights. We contrast the OOD detection performance of $\mathcal{S}$NN with these dropout strategies in Table 7 (Appendix). Notably, $\mathcal{S}$NN improves FPR95 by **37.90%** and **20.14%** as compared to Targeted Dropout (Gomez et al. 2019) and Adaptive Dropout (Ba and Frey 2013), respectively. Lastly, Wong, Santurkar, and Madry (2021) proposed an elastic net formulation to enforce sparsity for model interpretability.

8 Related Works

Training-based OOD detection methods. Commonly used approaches in this direction design training-time loss functions and regularization for better ID/OOD separability. De-Vries and Taylor (2017) proposed to augment the network with a confidence estimation branch. MOS (Huang and Li 2021) improves OOD detection by modifying the loss to use a pre-defined group structure. Another branch of studies based on train-time regularization have also shown promising and significant improvement in OOD detection performance (Lee et al. 2018a; Geifman and El-Yaniv 2019; Liu et al. 2020; Wei et al. 2022; Katz-Samuels et al. 2022; Du et al. 2022a,b; Ming, Fan, and Li 2022; Ming et al. 2023; Huang et al. 2023; Wang et al. 2023b; Bai et al. 2023; Wang et al. 2023a; Du et al. 2023; Zheng et al. 2023). Common approaches include training models to give predictions with lower confidences (Lee et al. 2018b; Hendrycks, Mazeika,

and Dietterich 2019; Wang et al. 2022) or higher energies (Liu et al. 2020; Du et al. 2022a; Song, Sebe, and Wang 2022) for outlier samples. Most regularization methods assume the availability of a large-scale and diverse outlier dataset which is not always realizable. Different from these methods, our proposed method does not require any additional outlier data. Rather, we formulate a novel train-time regularization approach based on learning feature subspaces. Our empirical analysis highlights the benefits of subspace learning for OOD detection.

Inference-based OOD detection methods. Studies in this domain mainly focus on designing scoring functions for OOD detection. These approaches include for example: (1) confidence-based methods (Hendrycks and Gimpel 2017; Liang, Li, and Srikant 2018), (2) energy-based methods (Liu et al. 2020; Morteza and Li 2022), (3) gradient-based method (Huang, Geng, and Li 2021), and (4) feature-based methods (Wilson et al. 2023). Some representative works include Mahalanobis distance (Lee et al. 2018b; Sehwag, Chiang, and Mittal 2021) and non-parametric KNN distance (Sun et al. 2022). However, the efficacy of these metrics is often limited in higher dimensions due to the "curse-of-dimensionality" (*c.f.* Section 4). Our paper targets precisely this critical yet underexplored problem. We show new insights that learning feature subspaces effectively alleviate this problem.

Subspace learning. To overcome limitations of "curse-of-dimensionality" in high dimensions, Ho (1998) explored the idea of random subspace for nearest neighbor. Subspace learning has also been used to improve search queries in high dimensions (Hund et al. 2015), feature selection (Liu and Motoda 2007), finding clusters in arbitrarily oriented spaces (Kriegel, Kröger, and Zimek 2009), and intrinsic dimensionality estimation (Houle et al. 2014). Different from these previous works, we explore class-relevant feature subspace learning for OOD detection.

9 Conclusion

Our work highlights the challenge of curse-of-dimensionality in OOD detection, and introduces a new solution called $\mathcal{S}$NN for detecting OOD samples. Traditional distance-based methods for OOD detection suffer from the curse-of-dimensionality, which makes it difficult to distinguish between ID and OOD samples in high-dimensional feature spaces. To address this issue, $\mathcal{S}$NN learns subspaces that capture the most informative feature dimensions for the task. Our method is supported by theoretical analysis, which shows that reducing the feature dimensions improves the distinguishability between ID and OOD samples. Extensive experiments demonstrate that $\mathcal{S}$NN achieves significant improvements in both OOD detection and ID calibration performance. We hope that our approach will inspire future research on this important problem.

Acknowledgements

References

Aggarwal, C. C.; Hinneburg, A.; and Keim, D. A. 2001. On the surprising behavior of distance metrics in high dimensional space. In *International conference on database theory*.

Ba, J.; and Frey, B. 2013. Adaptive dropout for training deep neural networks. In *Advances in neural information processing systems*.

Bai, H.; Canal, G.; Du, X.; Kwon, J.; Nowak, R. D.; and Li, Y. 2023. Feed Two Birds with One Scone: Exploiting Wild Data for Both Out-of-Distribution Generalization and Detection. In *Proceedings of the International Conference on Machine Learning*.

Beyer, K.; Goldstein, J.; Ramakrishnan, R.; and Shaft, U. 1999. When is "nearest neighbor" meaningful? In *International conference on database theory*.

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

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

DeVries, T.; and Taylor, G. W. 2017. Improved regularization of convolutional neural networks with cutout. In *arXiv preprint arXiv:1708.04552*.

Du, X.; Sun, Y.; Zhu, X.; and Li, Y. 2023. Dream the Impossible: Outlier Imagination with Diffusion Models. In *Advances in Neural Information Processing Systems*.

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

Du, X.; Wang, Z.; Cai, M.; and Li, Y. 2022b. VOS: Learning What You Don't Know by Virtual Outlier Synthesis. In *Proceedings of the International Conference on Learning Representations*.

Geifman, Y.; and El-Yaniv, R. 2019. Selectivenet: A deep neural network with an integrated reject option. In *Proceedings of the International Conference on Machine Learning*.

Gomez, A. N.; Zhang, I.; Kamalakara, S. R.; Madaan, D.; Swersky, K.; Gal, Y.; and Hinton, G. E. 2019. Learning sparse networks using targeted dropout. In *arXiv preprint arXiv:1905.13678*.

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

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

Hinneburg, A.; Aggarwal, C. C.; and Keim, D. A. 2000. What is the nearest neighbor in high dimensional spaces? In *26th Internat. Conference on Very Large Databases*.

Hinton, G. E.; Srivastava, N.; Krizhevsky, A.; Sutskever, I.; and Salakhutdinov, R. R. 2012. Improving neural networks by preventing co-adaptation of feature detectors. In *arXiv preprint arXiv:1207.0580*.

Ho, T. K. 1998. Nearest neighbors in random subspaces. In *Joint IAPR international workshops on statistical techniques in pattern recognition (SPR) and structural and syntactic pattern recognition (SSPR)*.

Houle, M. E.; Kriegel, H.-P.; Kröger, P.; Schubert, E.; and Zimek, A. 2010. Can shared-neighbor distances defeat the curse of dimensionality? In *International conference on scientific and statistical database management*.

Houle, M. E.; Ma, X.; Oria, V.; and Sun, J. 2014. Efficient algorithms for similarity search in axis-aligned subspaces. In *International Conference on Similarity Search and Applications*.

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

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

Huang, R.; and Li, Y. 2021. MOS: Towards Scaling Out-of-distribution Detection for Large Semantic Space. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*.

Huang, Z.; Xia, X.; Shen, L.; Han, B.; Gong, M.; Gong, C.; and Liu, T. 2023. Harnessing Out-Of-Distribution Examples via Augmenting Content and Style. In *Proceedings of the International Conference on Learning Representations*.

Hund, M.; Behrisch, M.; Färber, I.; Sedlmair, M.; Schreck, T.; Seidl, T.; and Keim, D. 2015. Subspace nearest neighbor search-problem statement, approaches, and discussion. In *International Conference on Similarity Search and Applications*.

Katz-Samuels, J.; Nakhleh, J.; Nowak, R.; and Li, Y. 2022. Training OOD Detectors in their Natural Habitats. In *Proceedings of the International Conference on Machine Learning*.

Kriegel, H.-P.; Kröger, P.; Schubert, E.; and Zimek, A. 2009. Outlier detection in axis-parallel subspaces of high dimensional data. In *Pacific-asia conference on knowledge discovery and data mining*.

Kriegel, H.-P.; Kröger, P.; and Zimek, A. 2009. Clustering high-dimensional data: A survey on subspace clustering, pattern-based clustering, and correlation clustering. *ACM Transactions on Knowledge Discovery from Data*.

Krizhevsky, A.; Hinton, G.; et al. 2009. Learning multiple layers of features from tiny images.

Lee, K.; Lee, H.; Lee, K.; and Shin, J. 2018a. Training Confidence-calibrated Classifiers for Detecting Out-of-Distribution Samples. In *Proceedings of the International Conference on Learning Representations*.

Lee, K.; Lee, K.; Lee, H.; and Shin, J. 2018b. A simple unified framework for detecting out-of-distribution samples and adversarial attacks. In *Advances in Neural Information Processing Systems*.

Liang, S.; Li, Y.; and Srikant, R. 2018. Enhancing The Reliability of Out-of-distribution Image Detection in Neural Networks. In *Proceedings of the International Conference on Learning Representations*.

Liu, H.; and Motoda, H. 2007. *Computational methods of feature selection*.

Liu, W.; Wang, X.; Owens, J.; and Li, Y. 2020. Energy-based Out-of-distribution Detection. In *Advances in Neural Information Processing Systems*.

Ming, Y.; Fan, Y.; and Li, Y. 2022. POEM: Out-of-Distribution Detection with Posterior Sampling. In *Proceedings of the International Conference on Machine Learning*.

Ming, Y.; Sun, Y.; Dia, O.; and Li, Y. 2023. How to Exploit Hyperspherical Embeddings for Out-of-Distribution Detection? In *Proceedings of the International Conference on Learning Representations*.

Morteza, P.; and Li, Y. 2022. Provable Guarantees for Understanding Out-of-distribution Detection. In *Proceedings of the AAAI Conference on Artificial Intelligence*.

Netzer, Y.; Wang, T.; Coates, A.; Bissacco, A.; Wu, B.; and Ng, A. Y. 2011. Reading Digits in Natural Images with Unsupervised Feature Learning.

Sehwag, V.; Chiang, M.; and Mittal, P. 2021. SSD: A Unified Framework for Self-Supervised Outlier Detection. In *Proceedings of the International Conference on Learning Representations*.

Song, Y.; Sebe, N.; and Wang, W. 2022. RankFeat: Rank-1 Feature Removal for Out-of-distribution Detection. In *Advances in Neural Information Processing Systems*.

Srivastava, N.; Hinton, G.; Krizhevsky, A.; Sutskever, I.; and Salakhutdinov, R. 2014. Dropout: a simple way to prevent neural networks from overfitting. *The journal of machine learning research*.

Sun, Y.; Guo, C.; and Li, Y. 2021. ReAct: Out-of-distribution Detection With Rectified Activations. In *Advances in Neural Information Processing Systems*.

Sun, Y.; and Li, Y. 2022. DICE: Leveraging Sparsification for Out-of-Distribution Detection. In *Proceedings of European Conference on Computer Vision*.

Sun, Y.; Ming, Y.; Zhu, X.; and Li, Y. 2022. Out-of-Distribution Detection with Deep Nearest Neighbors. In *Proceedings of the International Conference on Machine Learning*.

Tack, J.; Mo, S.; Jeong, J.; and Shin, J. 2020. Csi: Novelty detection via contrastive learning on distributionally shifted instances. In *Advances in neural information processing systems*.

Wang, Q.; Fang, Z.; Zhang, Y.; Liu, F.; Li, Y.; and Han, B. 2023a. Learning to Augment Distributions for Out-of-Distribution Detection. In *Advances in Neural Information Processing Systems*.

Wang, Q.; Ye, J.; Liu, F.; Dai, Q.; Kalander, M.; Liu, T.; HAO, J.; and Han, B. 2023b. Out-of-distribution Detection with Implicit Outlier Transformation. In *Proceedings of the International Conference on Learning Representations*.

Wang, Y.; Zou, J.; Lin, J.; Ling, Q.; Pan, Y.; Yao, T.; and Mei, T. 2022. Out-of-Distribution Detection via Conditional Kernel Independence Model. In *Advances in Neural Information Processing Systems*.

Wei, H.; Xie, R.; Cheng, H.; Feng, L.; An, B.; and Li, Y. 2022. Mitigating Neural Network Overconfidence with Logit Normalization. In *Proceedings of the International Conference on Machine Learning*.

Wilson, S.; Fischer, T.; Dayoub, F.; Miller, D.; and Sünderhauf, N. 2023. SAFE: Sensitivity-Aware Features for Out-of-Distribution Object Detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*.

Wong, E.; Santurkar, S.; and Madry, A. 2021. Leveraging sparse linear layers for debuggable deep networks. In *Proceedings of the International Conference on Machine Learning*.

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

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

Zhao, P.; and Lai, L. 2022. Analysis of knn density estimation. *IEEE Transactions on Information Theory*.

Zheng, H.; Wang, Q.; Fang, Z.; Xia, X.; Liu, F.; Liu, T.; and Han, B. 2023. Out-of-distribution Detection Learning with Unreliable Out-of-distribution Sources. In *Advances in Neural Information Processing Systems*.

Zhou, B.; Lapedriza, A.; Khosla, A.; Oliva, A.; and Torralba, A. 2017. Places: A 10 million image database for scene recognition. *IEEE transactions on pattern analysis and machine intelligence*.