Nearest Neighbor Guidance for Out-of-Distribution Detection

Jaewoo Park[1,2*]   Yoon Gyo Jung[3*]   Andrew Beng Jin Teoh[1†]
[1]Yonsei University   [2]AiV Co.   [3]Northeastern University

Abstract

Detecting out-of-distribution (OOD) samples are crucial for machine learning models deployed in open-world environments. Classifier-based scores are a standard approach for OOD detection due to their fine-grained detection capability. However, these scores often suffer from overconfidence issues, misclassifying OOD samples distant from the in-distribution region. To address this challenge, we propose a method called Nearest Neighbor Guidance (NNGuide) that guides the classifier-based score to respect the boundary geometry of the data manifold. NNGuide reduces the overconfidence of OOD samples while preserving the fine-grained capability of the classifier-based score. We conduct extensive experiments on ImageNet OOD detection benchmarks under diverse settings, including a scenario where the ID data undergoes natural distribution shift. Our results demonstrate that NNGuide provides a significant performance improvement on the base detection scores, achieving state-of-the-art results on both AUROC, FPR95, and AUPR metrics.

1. Introduction

The open-world environment poses a challenge for classification models as they may encounter input samples with unknown class labels, i.e., out-of-distribution (OOD) instances [44, 26, 42, 2, 4, 16]. The detection of such anomalous examples is crucial for preventing classifier malfunctions and potential harm. As a result, in safety-critical applications like self-driving [29, 5, 40, 19] and biosynthesis [46, 37], the OOD detection task plays a critical role in ensuring the dependable deployment of machine learning models. Therefore, a significant body of research has been dedicated to OOD detection [45].

The standard approach for OOD detection is to derive a score function from the trained network, such that the in-distribution (ID) samples exhibit relatively higher scores than OOD. One major paradigm in designing the detection

---

*Equal contribution. † Corresponding author: Andrew Beng Jin Teoh
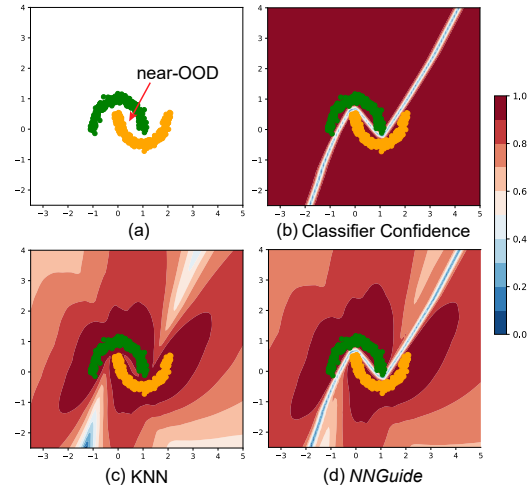

Figure 1.   (a) Out-of-distribution (OOD) instances can occur in any white region, including the small area between the in-distribution (ID) classes. For instance, given 'cat' and 'dog' as ID classes, images of 'fox' will be OOD instances near the ID data. (b) The classifier-based detector (*i.e.* confidence) assigns low scores on the small in-between area but suffers overconfidence issues. (c) Based on the distance information, KNN bounds the detection score on far-OOD regions. However, KNN lacks the fine-grained detection capability, and thus fails to detect the near-OOD. (d) NNGuide addresses both of these issues, reducing overconfidence in the far-OOD regions while achieving fine-grained detection.

score is to derive the score function based on the classifier's output signals, known as '*confidence*'. Examples of classifier-based scores include maximum softmax probability [15] and energy function [24]. A major advantage of the classifier-based detection scores is their ability to fully utilize the class-dependent information of ID data and provide fine-grained detection capability. However, the classifier-based scores may suffer from overconfidence in far OOD samples, limiting their effectiveness [10, 13].

In contrast, distance-based approaches (*e.g.* nearest neighbors [35] and Mahalanobis distance [22, 31]) detect OOD instances based on their distance to the ID data in the feature space. These approaches can certify low scores for far OOD regions but may not fully utilize class-dependent information, resulting in limited fine-grained detection ca-

pability.

Our work introduces *Nearest Neighbor Guidance* (NNGuide), a novel approach to improving classifier-based OOD detection scores and mitigating the issue of overconfidence. NNGuide achieves this by guiding the classifier confidence of a test input based on its similarity to its nearest neighbors in the ID bank set. As a result, NNGuide reduces the detection score in far-OOD regions while maintaining fine-grained detection capabilities.

In the toy experiment shown in Fig. 1, the classifier-based score demonstrates its ability to assign low detection scores to samples located in the small intermediate region between the two ID classes, which is where near-OOD instances may occur. However, the score exhibits excessively high values in the outer region of the ID data. On the other hand, the distance-based score provided by KNN is effective at assigning low detection scores to far-OOD samples, but it fails to credit low scores on the small intermediate region, indicating a lack of fine-grainiess. Our proposed detection score, NNGuide, mitigates the drawbacks of both methods; NNGuide assigns bounded low values for far-OOD samples while retaining fine-grained detection capability.

We conduct an extensive evaluation of NNGuide in the large-scale ImageNet-1k benchmark [18] across a variety of deep classification networks, achieving state-of-the-art results. Furthermore, we investigate the robustness of NNGuide by testing it on the ImageNet-1k-V2 dataset [36, 28], where the ID data undergoes natural distributional shifts. The presence of distribution shifts can lead to misidentifying ID samples as OOD and hence represents a challenging, realistic scenario. The final part of our experiments involves an extensive ablation analysis, where we investigate the key contributing factors to the effectiveness of NNGuide, as well as its compatibility with a broad range of classifier-based scores.

Contributions   The contributions of our work are summarized as follows:

- We propose a novel method called Nearest Neighbor Guidance (NNGuide) that guides the classifier-based detection score to reduce overconfidence in far-OOD regions while retaining its fine-grained detection capability.

- We attain state-of-the-art results on the ImageNet-1k OOD detection benchmarks and demonstrate the robustness of NNGuide by considering a challenging and realistic scenario where the ID ImageNet data undergoes a natural distributional shift.

- We provide an extensive and detailed ablation study, demonstrating the generality of NNGuide to a broad range of classifier-based scores.

We note that NNGuide is a post-hoc training-free inference method, and it is applicable to any standard deep classification networks.

2. Related Works

The OOD detection research primarily falls into two categories: network truncation [23, 33] and the design of a scalar score function to separate OOD instances from ID samples [15, 24].

Network truncation [23, 33] aims to increase the gap between ID and OOD samples by rectifying the propagated signals or weights of the network. For example, ODIN [23] perturbs the input signal using a gradient vector to increase the detection score, while ReAct [33] clips the hidden layer activation signals using a threshold. The clipped signals in ReAct are severely perturbed for OOD instances while being fairly retained for ID samples. Other methods, such as DICE, RankFeat, and BATS, follow the same principle as ReAct but rectify other types of signals. DICE [34] sparsifies the classification layer by removing fewer contributing weights therein. RankFeat [32] subtracts the rank-1 approximation of the feature map from the initial feature map. BATS [48] cut-outs signals that deviate from the batch norm statistics. The network truncation, however, cannot be used independently and need to be combined with a score function to detect OOD instances.

Another approach involves developing scalar score functions. These detection scores can be broadly categorized into two types: classifier-based and distance-based. The classifier-based scores, often referred to as *confidence*, leverage the classification layer of a neural network to derive the score. For example, [15] evaluated the effectiveness of the maximum output of softmax classifier probability (MSP). [24] proposed the energy function that can be viewed as a class conditional probability without bias. The maximum of logit [39, 14] captures both the class likelihood and the feature magnitude [8], and is shown to outperform the MSP counterpart. To utilize the class-dependent information extensively, [14] utilized the Kullback–Leibler (KL) divergence between the prediction and uniform distribution. GradNorm [17] on the other hand uses the norm of the gradient to minimize the KL score.

The other type of score-based approach is distance-based detectors. They identify an input sample as OOD based on its distance to the ID dataset in the feature space. One example is the Mahalanobis detector [22], which measures the minimum distance to the class-wise means based on the shared data feature covariance. A unified approach SSD [31] operates on the same principle as Mahalanobis but instead assumes that the ID samples follow a single Gaussian distribution with a single mean. In contrast, KNN [35] is non-parametric and therefore provides a more accurate representation of the distance to the boundary of the data man-

Algorithm 1 NNGuide Pseudocode, PyTorch-like

```python
def nnguide(z, s, Z, S, k):
    # z: 1-by-d array of a test sample feature
    # s: a scalar base confidence of the test sample
    # Z: n-by-d array of d-dimensional features of n
    #    samples from the bank set
    # S: n-by-1 array of base confidences of n
    #    samples from the bank set
    # k: the number of nearest neighbors
    Z = normalize(Z, dim=1)
    z = normalize(z, dim=1)

    g_topk, _ = matmul(z, (S*Z)).topk(k, dim=1)
    g = g_topk.mean(dim=1) # the guidance term

    return s*g # the guided score for the test sample
```

ifold. CIDER [25] shows that KNN particularly well fits to the networks with strong discriminative nature.

Classifier-based detectors exhibit low confidence scores on the class decision boundaries, and hence they are able to detect near-OOD instances around these boundaries. However, the classifier confidence is cursed to be overly confident in the far-OOD region [13]. Distance-based detectors on the other hand can certify low scores on the far-OOD regions. Nevertheless, they may struggle to assign low scores to samples located in the intermediate regions around ID classes, failing to detect near-OOD instances. This limitation can be particularly problematic for parametric methods like Mahalanobis and SSD when the modeled distribution does not align well with the true data manifold.

3. Preliminaries

The out-of-distribution (OOD) detection is formulated as follows: Let $\mathcal{X}$ denote the input space with the output space $\mathcal{Y} = \{1, \ldots, K\}$, where K is the number of classes in the in-distribution (ID) dataset. Let $f : \mathcal{X} \to \mathbb{R}^{|\mathcal{Y}|}$ be a neural network that outputs classification logits. The objective of OOD detection is to devise a detection score function S that determines whether a given test input $\mathbf{x} \in \mathcal{X}$ belongs to ID or OOD based on the score value $S(\mathbf{x})$:

$$\mathbf{x} \in \begin{cases} \text{ID} & \text{if} \quad S(\mathbf{x}) \geq \tau \\ \text{OOD} & \text{if} \quad S(\mathbf{x}) < \tau \end{cases} \tag{1}$$

The score function S is either derived from the classifier outputs $f(\mathbf{x})$ or by computing the distances to the hidden layer features $\phi(\mathbf{x})$ of the network f.

Terminology For brevity, we call the classifier-based detection score 'confidence'.

4. Method

4.1. Proposed method: NNGuide

Let $S_{base} : \mathcal{X} \to [0, \infty)$ be a given base confidence score function. We guide this base confidence by

$$S_{NNGuide}(\mathbf{x}) = S_{base}(\mathbf{x}) \cdot G(\mathbf{x}). \tag{2}$$

The *guidance term* $G(\mathbf{x})$ is derived from the ID nearest neighbors. Particularly, let $\{\mathbf{z}_1, \ldots, \mathbf{z}_n\}$ be a small bank set where n is a $\alpha\%$ of features randomly sampled from the train set. $\mathbf{z}_i = \phi(\mathbf{x}_i)$ is the feature computed from a bank set sample $\mathbf{x}_i$ by the penultimate layer of the network f. Let $s_i = S_{base}(\mathbf{x}_i)$ denote the confidence scores of $\mathbf{x}_i$. Then for a test input $\mathbf{x}$, the guidance term $G(\mathbf{x})$ is given by the average similarity to the k-nearest neighbors in the bank set

$$G(\mathbf{x}) = \frac{1}{k} \sum_{i=1}^{k} s_{(i)} \operatorname{sim}(\mathbf{z}_{(i)}, \mathbf{z}) \tag{3}$$

where $\mathbf{z} = \phi(\mathbf{x})$ is the test input feature, and sim is the cosine similarity. The reordered index (i) is given in the descending order of *confidence-scaled* nearest neighbor similarities

$$s_{(1)} \operatorname{sim}(\mathbf{z}_{(1)}, \mathbf{z}) \geq \cdots \geq s_{(n)} \operatorname{sim}(\mathbf{z}_{(n)}, \mathbf{z}). \tag{4}$$

Due to the confidence scale term $s_i = S_{base}(\mathbf{x}_i)$, the nearest neighbors are *selected in the high-confidence region*. This can enhance the utilization of more salient ID features while reducing the effect of possible outliers in ID. Overall, the confidence-scaled search makes the guided score more robust than the conventional KNN. The Pytorch-like pseudo algorithm is given in Algorithm 1.

Unless specified otherwise, we use the (negative) energy function $S_{base}(\mathbf{x}) = -\text{Energy}(\mathbf{x})$ as the default base confidence score due to its generality [24].

Theoretical Understanding Let $S = S_{NNguide}$, $\widehat{\phi}(\mathbf{x}) = \phi(\mathbf{x})/\|\phi(\mathbf{x})\|_2$, $\widehat{\mathbf{z}} = \mathbf{z}/\|\mathbf{z}\|_2$.

Proposition 1. *If* $\min_i \|\widehat{\phi}(\mathbf{x}) - \widehat{\mathbf{z}}_i\|_2 \geq 2$*, then* $S(\mathbf{x}) \leq 0$*. If* $\|\widehat{\phi}(\mathbf{x}) - \widehat{\mathbf{z}}_{(k)}\|_2 < \epsilon$*, then* $S(\mathbf{x}) > M(S_{base}(\mathbf{x}) - \epsilon/2)$ *if* $\min_{i \leq k} s_{(i)} > M$*, and* $S(\mathbf{x}) \leq \delta S_{base}(\mathbf{x})$ *if* $\max_{i \leq k} s_{(i)} \leq \delta$*.*

The proposition states that if a test sample $\mathbf{x}$ is far-distanced from the ID bank set in the feature space, then the guided score is certified to be low. On the other hand, if $\mathbf{x}$ is near to the ID, then the guidance up-scales the base confidence in the high-confidence region, while relatively down-scaling the base confidence around the low-confidence region (*e.g.* class decision boundaries). Thus, near the ID region, the guidance either retains or improves the fine-grained detection capability.

Training scheme	From scratch						Transfer learning					
Model	ResNet-50			MobileNet			ViT-B/16			RegNet-Y/16GF		
ID accuracy↑	78.73			72.15			85.3			86.01		
Detection method	FPR95↓	AUROC↑	AUPR↑	FPR95↓	AUROC↑	AUPR↑	FPR95↓	AUROC↑	AUPR↑	FPR95↓	AUROC↑	AUPR↑
MSP (ICLR'17[15])	49.54	87.44	96.71	77.04	79.46	94.65	48.74	87.67	96.94	43.37	88.95	97.30
MaxLogit (ICML'22[14])	42.12	90.49	97.57	78.06	77.99	94.10	37.62	89.29	97.11	26.00	92.91	98.16
KL (ICML'22[14])	40.01	90.80	97.63	88.70	68.85	91.56	38.44	89.01	97.03	24.74	93.07	98.18
ViM (CVPR'22[41])	29.86	**93.00**	**98.13**	76.27	73.59	92.53	35.16	91.04	97.73	21.39	94.89	98.74
Mahalanobis (NeurIPS'22[22])	44.58	90.93	97.76	**64.87**	80.29	94.56	39.34	91.73	98.08	32.15	93.07	98.36
SSD (ICLR'21[31])	40.94	91.47	97.82	77.78	69.01	90.18	59.74	80.38	94.74	40.64	90.22	97.57
GradNorm (NeurIPS'22[17])	28.92	93.00	98.09	89.88	65.28	89.88	38.04	89.28	96.91	82.86	62.98	87.76
KNN (ICML'22[35])	42.73	90.19	97.44	74.24	75.22	93.14	54.45	87.62	96.93	31.26	91.96	97.91
Energy (NeurIPS[24])	40.01	90.80	97.63	88.70	68.85	91.56	38.44	89.01	97.03	24.73	93.08	98.18
NNGuide (Ours)	**27.81**	92.89	98.03	65.92	**81.26**	**94.94**	34.20	92.14	98.10	**16.53**	**95.89**	**98.98**

Table 1. Results on ImageNet-1k. We report the average performance across five different OODs (*i.e*. iNaturalist, SUN, Places, Textures, OpenImage-O).

Detection method	Backbone	Venue	FPR95	AUROC	AUPR
ODIN*	ResNet-50	ICLR'18	56.48	85.41	-
GODIN*	ResNet-50	CVPR'20	66.07	82.02	-
DICE*	ResNet-50	ECCV'22	34.75	90.77	-
ReAct + DICE*	ResNet-50	ECCV'22	27.25	93.40	-
RankFeat*	ResNet-101	NeurIPS'22	36.80	92.15	-
BATS*	ResNet-50	NeurIPS'22	27.11	94.28	-
ASH*	ResNet-50	Arxiv'22	22.73	95.06	-
ReAct (+ Energy)*	ResNet-50	NeurIPS'21	31.43	92.95	-
ReAct + MSP	ResNet-50	reproduced	55.72	87.27	97.25
ReAct + MaxLogit	ResNet-50	reproduced	39.97	91.80	98.29
ReAct + KL	ResNet-50	reproduced	32.69	93.07	98.54
ReAct + ViM	ResNet-50	reproduced	26.06	94.83	98.85
ReAct + Mahalanobis	ResNet-50	reproduced	47.90	88.27	97.26
ReAct + SSD	ResNet-50	reproduced	56.17	83.77	95.91
ReAct + GradNorm	ResNet-50	reproduced	25.13	94.22	98.72
ReAct + KNN	ResNet-50	reproduced	42.42	89.46	97.46
ReAct + Energy	ResNet-50	reproduced	32.69	93.07	98.54
ReAct + NNGuide	ResNet-50	reproduced	**19.72**	**95.45**	**98.98**

Table 2. The comparison with the state-of-the-art network truncation methods. The average performance across four different OODs (*i.e*. iNaturalist, SUN, Places, and Textures) is reported. * indicates that the results are taken from the references.

5. Experiments

Our experiments on NNGuide are divided into the following parts: (1) We evaluate the performance of NNGuide on the standard ImageNet-1k OOD detection benchmark. (2) We examine the robustness of NNGuide against distribution shift. In this setting, the train ID data is ImageNet-1k while the test ID data is ImageNet-1k-V2 which comprises natural distribution shift examples. (3) NNGuide is evaluated on the small-scale CIFAR-100 [21] benchmark. (4) We conduct a thorough ablation study on NNGuide to identify its key components, assess its compatibility with other classifier-based scores, and determine the optimal conditions for its use. **Supplementary Sec. B provides complete experimental results.**

Configuration NNGuide involves two hyperparameters related to the k-nearest neighbor search, *i.e*. the number k of nearest neighbors, and sampling ratio $\alpha\%$ to construct the bank set from the train data. In all evaluations below, we follow the guideline of [35], and use $\alpha=1\%$ and $k=10$ to keep the balance between efficiency and performance. Extensive analysis of the hyperparameters is deferred to the ablation study in Sec. 5.4.3.

Comments on the computation speed The computational speed of nearest neighbor search has been extensively analyzed in [35] for OOD detection. [35] reports that KNN is as fast as or faster than most of the other detection methods in modern hardware and optimized libraries (*e.g*. faiss). NNGuide adds no computation overhead on the nearest neighbor search algorithm.

Evaluation metrics We evaluate OOD detection methods by the widely-used metrics: the false positive rate (FPR95) when the true positive rate of ID samples is at 95%, the area under the receiver operating characteristic curve (AUROC), and the area under the precision-recall curve (AUPR). In all the metrics, we regard the ID samples as positive. In addition, we report the closed-set classification accuracy of the model on the ID dataset.

5.1. Evaluation on the ImageNet-1k

Datasets In this evaluation, the train and test ID sets are all from ImageNet-1k [7]. For extensiveness, the detection method is evaluated on a diverse set of OOD datasets [18]: iNaturalist [38], SUN [43], Places [47], Textures [6], and OpenImage-O [41]. The OOD sets have no overlapping categories with ImageNet-1k. Though there is no strict criterion to differentiate between near-OOD and far-OOD [10], [44] categorizes iNaturalist and OpenImage-O as near-OOD and Textures as far-OOD. The other two OOD sets SUN and Places have overlapping characteristics. Overall, the five OOD sets involve diverse class semantics [18]. Hence, the average performance over these OOD sets indicates the robustness of the detection method against general OOD.

Backbone models We evaluate our proposed detection score NNGuide across four different model architectures ResNet-50 [12], MobileNet [30], ViT [9], and RegNet [27].

Training scheme	From scratch						Transfer learning					
Model	ResNet-50			MobileNet			ViT			RegNet		
ID accuracy↑	74.43			67.78			81.16			82.79		
Detection score	FPR95↓	AUROC↑	AUPR↑	FPR95↓	AUROC↑	AUPR↑	FPR95↓	AUROC↑	AUPR↑	FPR95↓	AUROC↑	AUPR↑
MSP	55.54	85.30	84.57	79.40	76.64	77.14	54.61	84.93	84.74	48.99	86.38	85.93
MaxLogit	47.97	88.54	88.01	80.19	75.17	75.06	45.97	86.32	84.75	34.17	89.67	87.97
KL	44.67	88.97	88.24	89.57	66.17	67.92	47.03	85.91	84.41	32.69	89.74	87.87
ViM	33.11	**91.88**	**90.75**	76.90	72.32	72.16	**39.13**	89.39	88.74	28.02	92.80	92.09
Mahalanobis	47.85	89.45	89.65	**65.27**	79.37	78.87	42.99	**90.29**	**90.76**	36.58	91.72	91.87
SSD	43.21	90.36	90.09	77.24	69.32	67.66	58.28	81.21	81.39	44.19	88.73	88.65
GradNorm	32.66	91.87	90.62	89.98	64.50	65.27	47.62	86.07	83.75	84.47	61.02	59.56
KNN	44.57	89.22	88.64	75.49	74.33	75.20	57.98	86.52	86.73	33.74	91.34	90.54
Energy	44.68	88.97	88.24	89.57	66.17	67.92	47.03	85.91	84.41	32.86	89.75	87.89
NNGuide	**30.78**	91.70	90.08	67.80	**79.40**	**78.93**	41.73	90.08	89.95	**21.97**	**94.17**	**93.44**

Table 3. Results on ImageNet-1k-V2. The average performance across five different OODs is reported.

ResNet-50 is a standard architecture for OOD evaluation, while MobileNet is a network designed particularly for efficiency. ViT partitions an image into multiple visual tokens and processes them by a deep stack of multi-head attention blocks, whose usage has shown excellent performance in language modeling. Unlike ViT, the RegNet architecture is targeted for both efficiency and performance. RegNet is constructed by applying a network search principle at the network-population level rather than a network level, making it robust across diverse environments including distribution shifts and domain generalization [3, 1].

All four models are trained on the training fold of ImageNet-1k, and the classification layers are strictly prohibited to see any instance from OOD datasets. The first two, ResNet-50 and MobileNet, are trained from scratch on the train set. The latter two, ViT and RegNet, on the other hand, are initialized from the pretrained weights on ImageNet-21k, and then the full weights are fine-tuned on ImageNet-1k. The particular versions we use are ViT-B/16 and RegNet-Y/16GF. The transfer learning scheme for ViT and RegNet is a more practical approach due to their higher ID (closed-set) accuracy and overall better detection performance.

5.1.1 Comparison on detection scores

We compare NNGuide with the state-of-the-art post-hoc detection score methods. The baselines include MSP, MaxLogit, ViM, Mahalanobis, SSD, GradNorm, KNN, and Energy detection scores, as described [Sec. 2]. Tab. 1 indicates that our proposed NNGuide is more effective than or on par with other detection scores across model architectures, training schemes, and evaluation metrics. Upon the RegNet model, particularly, we achieve a new state-of-the-art performance, by significantly outperforming all other methods. This shows that with a well-trained model, NNGuide can achieve robust OOD detection on the large-scale benchmark.

5.1.2 State-of-the-art performance with network truncator

The recent works in OOD detection showed that the truncation methods that rectify the hidden/output layer signals give excellent performance. The insight of these approaches is that a particular truncation function perturbs only the signals from the OOD instance while retaining those of ID samples, hence enhancing the score gap between ID and OOD. The truncators however cannot be used alone and require an external OOD detection score (such as MSP, Energy, or KNN).

For a fair comparison with truncation approaches, we combine our detection score NNGuide with ReAct. Tab. 2 shows that when combined with the simple truncator i.e. ReAct, our proposed NNGuide performs the best over other recently proposed network truncators in both the FPR95 and AUROC metrics. Particularly, NNGuide outperforms BATS, which is limited to the networks with batch normalization. 'ReAct + NNGuide' performs significantly better than a classifier rectifier DICE, even when used in conjunction with ReAct. In addition, When comparing ReAct combined with various detection scores, NNGuide demonstrates greater effectiveness and relevance than the other detection scores.

5.2. Evaluation against natural distribution shift

Datasets and configuration We evaluate the robustness of NNGuide against the natural distribution shift of the ID dataset. To this end, we consider the setting where the train ID data is ImageNet-1k and the test ID dataset is ImageNet-1k-V2 which consists of natural distribution shift samples. We note that both train and test ID datasets share the same semantics classes (*i.e.*, the 1k number of classes in ImageNet). OOD detection in this setup can be challenging since the detection score may incorrectly identify the test ID sample as an OOD instance due to the distribution difference between the test and train ID sets. As to model configurations, we use the same models that are used for the ImageNet-1k evaluation in Sec. 5.1.
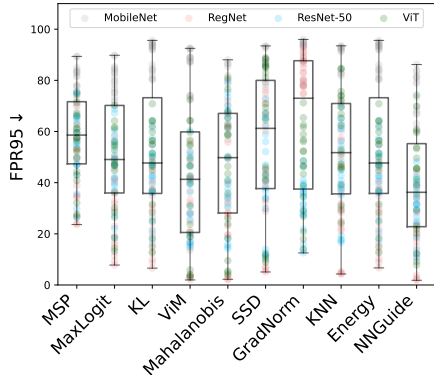
Figure 2. The performance summary of NNGuide and other detection scores on the ImageNet benchmarks across different models, test IDs, and OODs.

Results on ImageNet-1k-V2 Tab. 3 indicates that NNGuide is comparable to the state-of-the-art detection method ViM and Mahalanobis under MobileNet and the vision transformer, and significantly outperforms all other detection scores with ResNet-50 and RegNet. Even across MobileNet and ViT, however, NNGuide is overall more robust than Mahalanobis and ViM, indicated by the smaller fluctuation in the performance metrics.

Both the Mahalanobis and ViM detectors are known to excel in the ViT-type architecture due to the Gaussian nature of the vision transformer embedding space [11, 20]. We note however that ViT is suboptimal in this task even with large-scale pretraining and a much larger number of network parameters and inference time; ResNet-50 trained from scratch achieves better overall performance metrics than ViT.

We found that the RegNet architecture built by the network-population level search principle is shown to be the best in our comparison across all metrics. Under this backbone, our proposed NNGuide outperforms other detection scores by a large margin in the FPR95 metric.

We present a summary of our performance evaluation in Fig. 2. The figure demonstrates that NNGuide exhibits less performance variation and better average performance compared to all other detection scores, including ViM, which is also based on fusing classifier signals with distance information.

5.3. Evaluation on the CIFAR-100 benchmark

We evaluate NNGuide on the small-scale CIFAR-100 by training ResNet-18 on the train fold from scratch. Each class of CIFAR-100 contains a small number of low-resolution images, and hence the trained model can be suboptimal for both classification and OOD detection [39]. Tab. 4 shows the average performance of NNGuide against five different OODs that are commonly used for evaluation

	FPR95↓	AUROC↑	AUPR↑
MSP	72.53	82.34	83.05
MaxLogit	67.92	85.14	85.61
KL	66.27	85.47	85.84
ViM	80.07	77.70	78.63
Mahalanobis	77.54	78.88	79.18
SSD	83.21	69.05	65.52
GradNorm	66.90	76.71	72.14
KNN	70.07	84.17	84.23
Energy	66.27	85.47	85.84
NNGuide	**64.56**	**86.39**	**86.96**

Table 4. Results on the CIFAR-100 benchmark with the ResNet-18 model trained on CIFAR-100 from scratch, achieving 75.66% ID accuracy. The average performance across five different OODs is reported.

	FPR95↓	AUROC↑	AUPR↑
baselines:			
KNN	42.73	90.19	97.44
KNN with average similarity	43.96	90.71	97.64
Energy	40.01	90.80	97.63
naive fusion:			
Product fusion	33.53	92.17	97.92
Sum fusion	34.03	92.27	97.96
Max fusion	42.72	90.19	97.44
Min fusion	40.01	90.80	97.63
missing core components			
Mahalanobis guidance	37.19	92.00	97.97
Guidance term only	30.31	91.58	97.42
W/O confidence scaling	36.23	91.81	97.87
NNGuide	**27.81**	**92.89**	**98.03**

Table 5. Ablation study on the components of NNGuide. The average performance across five different OODs is reported.

(*i.e.* CIFAR-10, SVHN, resized LSUN, resized ImageNet, and iSUN). NNGuide outperforms other baseline detection methods. However, unlike the ImageNet benchmarks, the performance boost by NNGuide is marginal on the CIFAR-100 evaluation protocol. This is due to the suboptimal model trained on low-quality ID data. Such a limitation of NNGuide is more carefully analyzed in the ablation study (Sec. 5.4.4).

5.4. Ablation

The ablation study is divided into several parts. (1) We examine the compatibility of NNGuide with other classifier-based detection scores besides the negative energy function. (2) We analyze which components of NNGuide contribute to improving the classifier-based confidence score. (3) We evaluate the impact of the hyperparameters k and α related to the nearest neighbor search. (4) We present a limitation analysis of NNGuide from various aspects, highlighting the necessary requirements for its optimal usage.
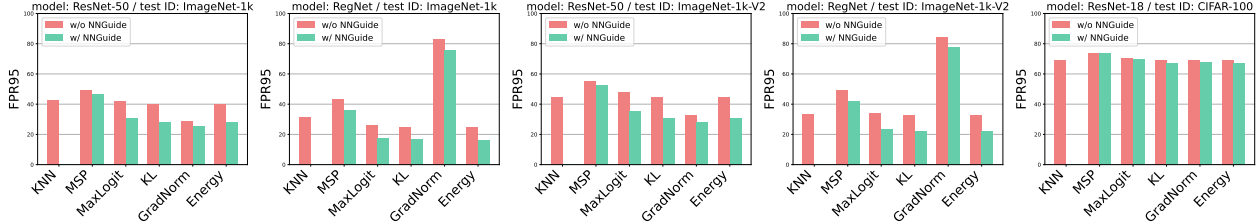
Figure 3. The compatibility to other classifier-based scores. The average performance across five different OODs is reported.
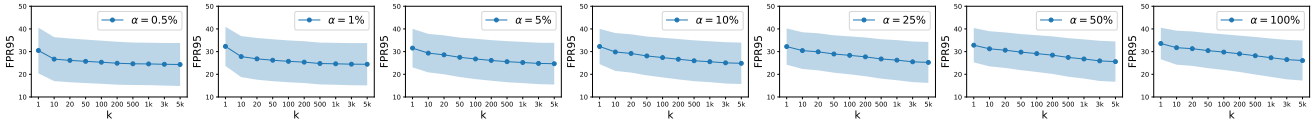

Figure 4. The hyperparameter analysis on the number k of nearest neighbors and the sampling ratio α.

5.4.1 Compatibility to other classifier-based confidence scores

We extend the use of NNGuide beyond the energy function by evaluating its compatibility with other classifier-based confidence scores, *i.e.*, MSP, MaxLogit, KL, and Grad-Norm. Fig. 3 demonstrates that NNGuide effectively enhances the performance of all considered scores. Notably, the performance improvement is remarkable, particularly for the scores that fully utilize class-dependent information (both target and non-target class outputs) without the softmax nonlinearity. However, we note that the final performance of NNGuide heavily relies on the base confidence score. As such, NNGuide may not be effective when the base score is poor. We further analyze the limitations of NNGuide in Sec. 5.4.4.

5.4.2 Ablation on the components of NNGuide

NNGuide consists of the base classifier confidence score $S_{base}(\mathbf{x})$ and the nearest-neighbor-based guidance term $G(\mathbf{x})$. The guidance term can be further broken down to the confidence scaling in Eq. (4) and the similarity ensemble in Eq. (3).

Tab. 5 shows the overall results of evaluations conducted to ablate each component of NNGuide. Here, the original KNN detector as formulated in [35] detects an OOD instance based on only the similarity to the k-th nearest neighbor; *i.e.* $S_{\text{KNN}}(x) = \text{sim}(\mathbf{z}_{(k)}, \mathbf{z})$. To see the effect of similarity ensemble, we modify the KNN detector to compute the average of top-k similarities to nearest ID instances; *i.e.* $S_{\text{avg-KNN}}(\mathbf{x}) = \sum_{i=1}^{k} \text{sim}(\mathbf{z}_{(k)}, \mathbf{z})$. As indicated by 'KNN with average similarity', the similarity ensemble alone does not boost the performance.

We argue that similarity ensemble is effective only when combined with confidence scaling. To validate this, we evaluate the performance of the guidance term $G(\mathbf{x})$. The term can be considered as a weighted KNN, where the weights are the confidences s_i. Indicated by 'Guidance term only', Tab. 5 shows a notable improvement to the original KNN. As discussed in Sec. 4, the nearest neighbor search based on *confidence-scaled* similarities selects the bank set instances in the high-confidence region. Hence, this search algorithm operates with the most salient ID features, ignoring possible outliers.

We further verify the effectiveness of the confidence-scaled nearest neighbor search by removing the confidence-scaling component from NNGuide. The result indicated by 'W/O confidence scaling' shows that confidence scaling is a significant factor in NNGuide.

We note that the Mahalanobis (density) score can also be used to bind the overconfidence of the classifier on the far-OOD region. Hence, we evaluate its impact by substituting the guiding term with the Mahalanobis distance. Tab. 5 shows that the guidance by Mahalnobis score is not as effective as NNGuide. The disadvantage of Mahalanobis may stem from a strong parametric assumption that the ID features should be Gaussian.

Finally, we test with other types of fusion techniques. We find that a naive combination of the KNN and classifier-based detection score by basic algebraic operations such as min, max, sum, and the product is not as effective as the NNGuide. Although these basic fusion approaches are aligned with our high-level objective, the confidence-scaled nearest neighbor search and similarity ensemble parts are missing therein. Hence, the naive fusion detectors could be neither fine-grained nor robust, testified by their worse performances.

5.4.3 Analysis of the hyperparameters

We evaluate the impact of hyperparameters α and k in NNGuide. We consider $\alpha \in \{0.5, 1, 5, 10, 25, 50, 100\}$ and $k \in \{1, 10, 20, 50, 100, 200, 500, 1000, 3000, 5000\}$ and for the sampling ratio and a number of neighbors, respec-
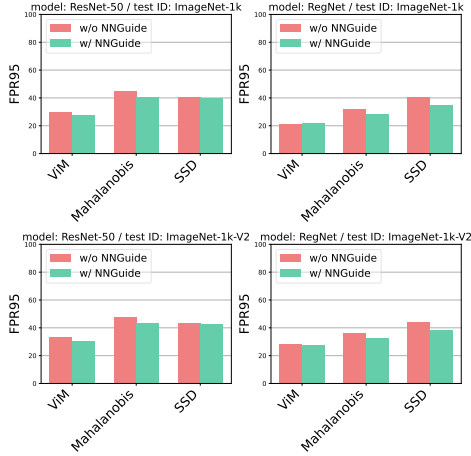
Figure 5. NNGuide could be suboptimal when combined with distance-based scores.
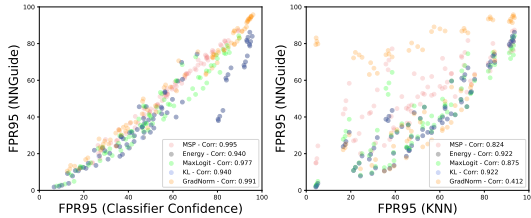

Figure 6. The dependency of NNGuide on the classifier confidence and KNN.

tively. Fig. 4 indicates that the performance of NNGuide is fairly robust across different k as long as $k \geq 10$. Moreover, the performance has a consistent trend across different sampling ratios α. This is in contrast to the vanilla KNN detection score; as reported in [35], the KNN score exhibits a degree of performance fluctuation under the variation of k and α. In addition, the performance variance is lower in the small sampling regime (*i.e.* small $\alpha\%$), suggesting that the hyperparameters of NNGuide can be fairly easily tuned.

5.4.4 Limitation analysis of NNGuide

Despite the superiority of NNGuide compared to other detection scores, it is not perfect. We analyze its limitation, figuring out necessary requirements for NNGuide.

The necessity of classifier for NNGuide With the nearest neighbor guidance by Eq. (2), the base confidence score is assumed to be a classifier-based score with fine-grained detection capability. One possible approach is to extend the base confidence score to distance-based scores (*e.g.* Mahalanobis, SSD, and ViM), which however may lack a fine-grained nature. We argue that NNGuide could be suboptimal if the base confidence is a distance-based score. Guiding a distance-based score by the nearest neighbor guidance may not significantly boost its detection capability. As

shown in Fig. 5, the improvement by NNGuide is inconsistent and sometimes marginal. The guidance is particularly insignificant when the base score is ViM, where the overly low energy in the far-OOD region is already mitigated by the orthogonal distance to the ID subspace.

In some cases, NNGuide improves Mahalanobis and SSD. We believe this is due to the following fact: Mahalanobis and SSD represent poor distance functions when the data feature has deviated from Gaussian. On the other hand, the nearest neighbors represent the data boundary more accurately, providing a better distance function. Thus, the guidance by nearest neighbors refines the Mahlanobis and SSD distances.

Dependency on both classifier confidence and KNN NNGuide is formulated by nearest neighbors and the classifier outputs. Hence, the performance of NNGuide inevitably depends on both KNN and the classifier's confidence. Fig. 6 indicates a strong linear correlation between NNGuide and the base classifier's confidence in terms of the performance metric. NNGuide exhibits a strong correlation with KNN only when the classifier-based score demonstrates robust OOD detection capability. Specifically, the correlation is not observed with suboptimal classifier confidences such as GradNorm and MSP (Fig. 3). Conversely, NNGuide shows a strong correlation with KNN when the confidences are based on Energy, KL, and MaxLogit. This correlation trend suggests that the optimal usage of NNGuide requires both good classifier confidence and a strong feature extractor.

The analysis also explains our results on ResNet-18 (CIFAR-100) and RegNet (ImageNet). The suboptimal ResNet-18 trained on CIFAR-100 from scratch likely produces poor classification outputs and extracted features. Thus, NNGuide improves the base score only marginally. In contrast, the RegNet model trained on large-scale datasets with the best principles attains a highly discriminative classifier and representation. Accordingly, in RegNet, the performance boost by NNGuide is significant.

6. Conclusion

We proposed a novel method for OOD detection called the nearest neighbor guidance (NNGuide) that improves a classifier's confidence by guiding it with the nearest neighbors' information. NNGuide prevents the overconfidence of the classifier while retaining its fine-grained detection capability, thereby achieving balanced robustness against both far- and near-OODs. NNGuide has been examined extensively on the large-scale ImageNet-1k benchmarks, including the natural distribution shift and transfer learning scenarios. NNGuide has shown to be robust across different model backbones and OODs, achieving state-of-the-art performance with RegNet.

References

[1] Devansh Arpit, Huan Wang, Yingbo Zhou, and Caiming Xiong. Ensemble of averages: Improving model selection and boosting performance in domain generalization. *arXiv preprint arXiv:2110.10832*, 2021. 5

[2] Wentao Bao, Qi Yu, and Yu Kong. Evidential deep learning for open set action recognition. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 13349–13358, 2021. 1

[3] Junbum Cha, Kyungjae Lee, Sungrae Park, and Sanghyuk Chun. Domain generalization by mutual-information regularization with pre-trained models. In *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XXIII*, pages 440–457. Springer, 2022. 5

[4] Robin Chan, Matthias Rottmann, and Hanno Gottschalk. Entropy maximization and meta classification for out-of-distribution detection in semantic segmentation. In *Proceedings of the ieee/cvf international conference on computer vision*, pages 5128–5137, 2021. 1

[5] Kashyap Chitta, Aditya Prakash, and Andreas Geiger. Neat: Neural attention fields for end-to-end autonomous driving. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 15793–15803, 2021. 1

[6] Mircea Cimpoi, Subhransu Maji, Iasonas Kokkinos, Sammy Mohamed, and Andrea Vedaldi. Describing textures in the wild. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3606–3613, 2014. 4

[7] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009. 4

[8] Akshay Raj Dhamija, Manuel Günther, and Terrance Boult. Reducing network agnostophobia. *Advances in Neural Information Processing Systems*, 31, 2018. 2

[9] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020. 4, 12

[10] Zhen Fang, Yixuan Li, Jie Lu, Jiahua Dong, Bo Han, and Feng Liu. Is out-of-distribution detection learnable? *arXiv preprint arXiv:2210.14707*, 2022. 1, 4

[11] Stanislav Fort, Jie Ren, and Balaji Lakshminarayanan. Exploring the limits of out-of-distribution detection. *Advances in Neural Information Processing Systems*, 34:7068–7081, 2021. 6

[12] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016. 4, 12

[13] Matthias Hein, Maksym Andriushchenko, and Julian Bitterwolf. Why relu networks yield high-confidence predictions far away from the training data and how to mitigate the problem. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 41–50, 2019. 1, 3

[14] Dan Hendrycks, Steven Basart, Mantas Mazeika, Mohammadreza Mostajabi, Jacob Steinhardt, and Dawn Song. Scaling out-of-distribution detection for real-world settings. *arXiv preprint arXiv:1911.11132*, 2019. 2, 4

[15] Dan Hendrycks and Kevin Gimpel. A baseline for detecting misclassified and out-of-distribution examples in neural networks. *arXiv preprint arXiv:1610.02136*, 2016. 1, 2, 4

[16] Dan Hendrycks, Kevin Zhao, Steven Basart, Jacob Steinhardt, and Dawn Song. Natural adversarial examples. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15262–15271, 2021. 1, 17

[17] Rui Huang, Andrew Geng, and Yixuan Li. On the importance of gradients for detecting distributional shifts in the wild. *Advances in Neural Information Processing Systems*, 34:677–689, 2021. 2, 4

[18] Rui Huang and Yixuan Li. Mos: Towards scaling out-of-distribution detection for large semantic space. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8710–8719, 2021. 2, 4

[19] Sanghun Jung, Jungsoo Lee, Daehoon Gwak, Sungha Choi, and Jaegul Choo. Standardized max logits: A simple yet effective approach for identifying unexpected road obstacles in urban-scene segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 15425–15434, 2021. 1

[20] Rajat Koner, Poulami Sinhamahapatra, Karsten Roscher, Stephan Günnemann, and Volker Tresp. Oodformer: Out-of-distribution detection transformer. *arXiv preprint arXiv:2107.08976*, 2021. 6

[21] Alex Krizhevsky, Geoffrey Hinton, et al. Learning multiple layers of features from tiny images. 2009. 4

[22] Kimin Lee, Kibok Lee, Honglak Lee, and Jinwoo Shin. A simple unified framework for detecting out-of-distribution samples and adversarial attacks. *Advances in neural information processing systems*, 31, 2018. 1, 2, 4

[23] Shiyu Liang, Yixuan Li, and Rayadurgam Srikant. Enhancing the reliability of out-of-distribution image detection in neural networks. *arXiv preprint arXiv:1706.02690*, 2017. 2

[24] Weitang Liu, Xiaoyun Wang, John Owens, and Yixuan Li. Energy-based out-of-distribution detection. *Advances in neural information processing systems*, 33:21464–21475, 2020. 1, 2, 3, 4

[25] Yifei Ming, Yiyou Sun, Ousmane Dia, and Yixuan Li. How to exploit hyperspherical embeddings for out-of-distribution detection? *arXiv preprint arXiv:2203.04450*, 2022. 3, 17

[26] Yuchun Pu, Zhenghui Feng, Zhonglei Wang, Zhenyu Yang, and Jianping Li. Anomaly detection for in situ marine plankton images. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3661–3671, 2021. 1

[27] Ilija Radosavovic, Raj Prateek Kosaraju, Ross Girshick, Kaiming He, and Piotr Dollár. Designing network design spaces. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10428–10436, 2020. 4, 12

[28] Benjamin Recht, Rebecca Roelofs, Ludwig Schmidt, and Vaishaal Shankar. Do imagenet classifiers generalize to imagenet? In *International conference on machine learning*, pages 5389–5400. PMLR, 2019. 2, 13

[29] Xuanchi Ren, Tao Yang, Li Erran Li, Alexandre Alahi, and Qifeng Chen. Safety-aware motion prediction with unseen vehicles for autonomous driving. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 15731–15740, 2021. 1

[30] Mark Sandler, Andrew Howard, Menglong Zhu, Andrey Zhmoginov, and Liang-Chieh Chen. Mobilenetv2: Inverted residuals and linear bottlenecks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4510–4520, 2018. 4, 12

[31] Vikash Sehwag, Mung Chiang, and Prateek Mittal. Ssd: A unified framework for self-supervised outlier detection. *arXiv preprint arXiv:2103.12051*, 2021. 1, 2, 4

[32] Yue Song, Nicu Sebe, and Wei Wang. Rankfeat: Rank-1 feature removal for out-of-distribution detection. *arXiv preprint arXiv:2209.08590*, 2022. 2

[33] Yiyou Sun, Chuan Guo, and Yixuan Li. React: Out-of-distribution detection with rectified activations. *Advances in Neural Information Processing Systems*, 34:144–157, 2021. 2

[34] Yiyou Sun and Yixuan Li. Dice: Leveraging sparsification for out-of-distribution detection. In *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XXIV*, pages 691–708. Springer, 2022. 2

[35] Yiyou Sun, Yifei Ming, Xiaojin Zhu, and Yixuan Li. Out-of-distribution detection with deep nearest neighbors. In *International Conference on Machine Learning*, pages 20827–20840. PMLR, 2022. 1, 2, 4, 7, 8

[36] Rohan Taori, Achal Dave, Vaishaal Shankar, Nicholas Carlini, Benjamin Recht, and Ludwig Schmidt. Measuring robustness to natural distribution shifts in image classification. *Advances in Neural Information Processing Systems*, 33:18583–18599, 2020. 2, 13

[37] Jessica Vamathevan, Dominic Clark, Paul Czodrowski, Ian Dunham, Edgardo Ferran, George Lee, Bin Li, Anant Madabhushi, Parantu Shah, Michaela Spitzer, et al. Applications of machine learning in drug discovery and development. *Nature reviews Drug discovery*, 18(6):463–477, 2019. 1

[38] Grant Van Horn, Oisin Mac Aodha, Yang Song, Yin Cui, Chen Sun, Alex Shepard, Hartwig Adam, Pietro Perona, and Serge Belongie. The inaturalist species classification and detection dataset. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 8769–8778, 2018. 4

[39] Sagar Vaze, Kai Han, Andrea Vedaldi, and Andrew Zisserman. Open-set recognition: A good closed-set classifier is all you need. *arXiv preprint arXiv:2110.06207*, 2021. 2, 6

[40] Tomas Vojir, Tomáš Šipka, Rahaf Aljundi, Nikolay Chumerin, Daniel Olmeda Reino, and Jiri Matas. Road anomaly detection by partial image reconstruction with segmentation coupling. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 15651–15660, 2021. 1

[41] Haoqi Wang, Zhizhong Li, Litong Feng, and Wayne Zhang. Vim: Out-of-distribution with virtual-logit matching. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4921–4930, 2022. 4

[42] Jhih-Ciang Wu, Ding-Jie Chen, Chiou-Shann Fuh, and Tyng-Luh Liu. Learning unsupervised metaformer for anomaly detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4369–4378, 2021. 1

[43] Jianxiong Xiao, James Hays, Krista A Ehinger, Aude Oliva, and Antonio Torralba. Sun database: Large-scale scene recognition from abbey to zoo. In *2010 IEEE computer society conference on computer vision and pattern recognition*, pages 3485–3492. IEEE, 2010. 4

[44] Jingkang Yang, Pengyun Wang, Dejian Zou, Zitang Zhou, Kunyuan Ding, Wenxuan Peng, Haoqi Wang, Guangyao Chen, Bo Li, Yiyou Sun, et al. Openood: Benchmarking generalized out-of-distribution detection. *arXiv preprint arXiv:2210.07242*, 2022. 1, 4

[45] Jingkang Yang, Kaiyang Zhou, Yixuan Li, and Ziwei Liu. Generalized out-of-distribution detection: A survey. *arXiv preprint arXiv:2110.11334*, 2021. 1

[46] Shuangjia Zheng, Tao Zeng, Chengtao Li, Binghong Chen, Connor W Coley, Yuedong Yang, and Ruibo Wu. Deep learning driven biosynthetic pathways navigation for natural products with bionavi-np. *Nature Communications*, 13(1):3342, 2022. 1

[47] Bolei Zhou, Agata Lapedriza, Aditya Khosla, Aude Oliva, and Antonio Torralba. Places: A 10 million image database for scene recognition. *IEEE transactions on pattern analysis and machine intelligence*, 40(6):1452–1464, 2017. 4

[48] Yao Zhu, YueFeng Chen, Chuanlong Xie, Xiaodan Li, Rong Zhang, Hui Xue, Xiang Tian, Yaowu Chen, et al. Boosting out-of-distribution detection with typical features. *arXiv preprint arXiv:2210.04200*, 2022. 2