ENCORE: A Neural Collapse Perspective on Out-of-Distribution Detection in Deep Neural Networks

A.Q.M. Sazzad Sayyed[†], Nathaniel D. Bastian[‡] and Francesco Restuccia[†]
[†]Northeastern University [‡]United States Military Academy
Corresponding author: sayyed.a@northeastern.edu

Abstract

Out-of-Distribution (OOD) detection is of paramount importance in guaranteeing safe and reliable deployment of a Deep Neural Network (DNN) model in real-world settings. However, most OOD detection approaches still lack motivation rooted in established properties of the DNNs. This disconnect between the proposed approach and theoretical underpinning to measurable DNN properties makes these approaches unreliable. To bridge this gap, we take a different perspective to using energy scoring for OOD detection. Specifically, we look at energy score through the lens of the properties of neural collapse and observe that simple feature scaling can improve the separation between In-Distribution (ID) and OOD inputs. Based on this observation, we propose ENCORE, which scales features of each input adaptively and uses them to obtain modified logits based on insights from theory of neural collapse. We show that ENCORE outperforms state-of-the-art approaches – for example, by 1.37% on CIFAR10 and by 1.07% on Imagenet benchmarks. We are sharing the implementation code[1].

1. Introduction

In real-world settings, a DNN may encounter corrupted inputs or entirely new classes not encountered during training. These are commonly referred to as OOD inputs. Existing OOD detection methods fall into two broad categories, namely *training-time regularization* and *post-hoc* approaches [25]. Training-time methods improve OOD robustness by modifying the training objective [10, 16, 22]. In contrast, post-hoc methods do not require any changes to training as they work with already trained models. Prior work has investigated confidence-based methods of OOD detection that rely on the DNN output probabilities [8, 9, 12, 13]. On the other hand, feature-based methods examine internal representations [3, 19, 21]. The key limitation is that existing approaches are still often empirical in nature and most do not connect with any established property of the DNNs. As a result, it becomes harder to pinpoint the conditions under which an OOD detector might fail.

Among post-hoc methods, energy-based scoring has become one of the most widely adopted approaches for OOD detection. It computes a scalar score from the DNN logits, typically using the log-sum-exp formulation [13]. The appeal of this method lies in its simplicity, effectiveness, and compatibility with pre-trained softmax classifiers. Since the energy score captures the probability of a single input to occur, it serves as a proxy for uncertainty. Lower energy values usually indicate more confident predictions on ID data. However, despite its success, the energy score is not always reliable as it can still assign low energy (i.e., high confidence) to OOD inputs, especially when the model is poorly calibrated or the logits are spuriously high. This overconfidence fundamentally limits existing energy-based methods.

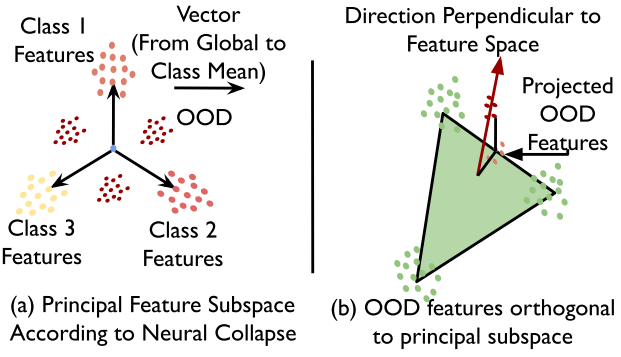

Figure 1. (a) With neural collapse, the ID features cluster around respective class means. As OOD inputs do not align with any class, increasing the feature vector increases the separation between ID and OOD inputs. (b) OOD inputs are orthogonal to ID feature space according to neural collapse. As such, the norm of the OOD inputs projected onto the feature space is significantly different than ID inputs.

Key Idea. In this work, we provide a new perspective on energy-scoring for OOD detection through the lens of *neural collapse* [17]. The latter is a phenomenon observed during the terminal phase of training of deep classifiers, where class features and weights form a highly structured symmet-

[1]https://github.com/Restuccia-Group/ENCORE.git

ric geometric configuration. Neural collapse offers a new perspective on how energy behaves for ID and OOD inputs. We leverage this new perspective (illustrated in Figure 1) to propose *Energy with Neural Collapse-Oriented Representation Embedding (ENCORE)*, an improved OOD score based on theoretical observations of neural collapse. Specifically, ENCORE computes the margin between the energy score of the ID and OOD inputs under the assumption of neural collapse. *Importantly, we show that this margin increases with the increase in feature norm, improving the OOD detection performance*. Next, we incorporate the deviations from the assumptions of the neural collapse and construct an adaptive feature-norm-based logit with which the energy score is computed. This approach minimally increases the complexity of the energy score while significantly improving the OOD detection performance, which performs competitively or better than the SOTA methods in both near-OOD and far-OOD scenarios. To summarize, our contributions are as follows:

- We provide a theoretical analysis of energy-based OOD detection under neural collapse, showing that the margin between ID and OOD energy scores increases with feature norm. Based on this insight, we propose an adaptive feature-norm-based energy score that accounts for deviations from ideal neural collapse;

- We validate our method through comprehensive experiments across diverse architectures and standard OOD benchmarks, demonstrating competitive or superior performance in both near-OOD and far-OOD settings without requiring retraining or architectural changes. In far-OOD settings, ENCORE achieves 1.37% better False Positive Rate (FPR) than the nearest state-of-the-art (SOTA) approach for CIFAR10 benchmark while it lags behind the SOTA by only 0.26% for Imagenet benchmark. For near-OOD, ENCORE outperforms SOTA by 1.07%, and 0.35% on Imagenet and CIFAR-10 benchmarks respectively. We also show that ENCORE performs consistently for both convolutional architecture (e.g., ResNet, ConvNext, RepVGG) as well as vision transformers (ViTs).

2. Background and Problem Formulation

Problem Formulation. Similar to prior works, we cast our problem in the context of supervised multi-class classification. We represent a DNN as function $\mathcal{F}_\theta(.)$ parameterized by θ. $\mathbf{X}$ represents the input random variable and $\mathbf{x}$ represents its particular realization. Probability distributions are denoted as $\mathcal{P}$, and their corresponding probability measures are denoted with $\mathbf{P}$. We use $\mathcal{D}_{id}$ to denote ID dataset and $\mathcal{P}_{id}$ for ID data distribution. We refer to $\mathcal{P}_{ood}$ to denote the OOD distribution, while $\mathcal{D}_{ood}$ denotes the OOD dataset.

We define the problem of OOD detection as a binary classification problem. Specifically, we aim to design a score function $S(\mathbf{x})$ such that, $S(\mathbf{x}) \sim \mathcal{P}_{id}^S$ when $\mathbf{x} \in \mathcal{D}_{id}$ and $S(\mathbf{x}) \sim \mathcal{P}_{ood}^S$ when $\mathbf{x} \in \mathcal{D}_{ood}$. Then the binary classification decision can be formulated as finding the indicator function $\mathbb{1}(\mathbf{x})$ such that

$$\mathbb{1}(\mathbf{x}) = \begin{cases} 1, & \text{if } S(\mathbf{x}) > \tau \\ 0, & \text{if } S(\mathbf{x}) < \tau. \end{cases} \tag{1}$$

A value of 1 in the indicator function implies $\mathbf{x} \sim \mathcal{P}_{id}$ and a value of 0 means $\mathbf{x} \sim \mathcal{P}_{ood}$. The threshold τ controls the performance of the binary classifier and trades off True Positive Rate (TPR) for lower FPR. Ideally, the distributions $\mathcal{P}_{id}^S$ and $\mathcal{P}_{ood}^S$ have zero overlap, and the threshold τ can be set such that 100% TPR can be achieved for 0% FPR. In practice, we aim to find the threshold that can achieve a certain TPR (typically 95%) and measure the performance of the classifier with the obtained FPR value. Notice that Eq. (1) implicitly assumes that the score for the ID samples is higher than the score for the OOD samples. Even if this is not the case, the score function can be inverted (multiplied with -1) to comply with Eq. (1).

Background on Neural Collapse (NC). This phenomenon describes the emergent geometric structure of the last-layer features and weights of a DNN when training drives the classification error to zero while continuing to minimize the cross-entropy loss. This phase, referred to as the *Terminal Phase of Training (TPT)*, results in convergence to a highly symmetric and interpretable form.

Let $\mathcal{F}_\theta(\cdot)$ be a DNN trained on an in-distribution dataset $\mathcal{D}_{id} = \{(\mathbf{x}_i, y_i)\}_{i=1}^N$, where $y_i \in \{1, 2, \ldots, C\}$ and C is the number of classes. Let $\phi_\theta(\cdot)$ denote the feature extractor such that the penultimate-layer feature representation is $\mathbf{h}_i = \phi_\theta(\mathbf{x}_i) \in \mathbb{R}^d$. Denote the class-conditional feature mean as $\boldsymbol{\mu}_c = \mathbb{E}_{\mathbf{x}\sim\mathcal{P}_{id}}[\phi_\theta(\mathbf{x}|\mathbf{y} = \mathbf{c})]$ and the global mean as $\boldsymbol{\mu}_G = \frac{1}{C}\sum_{c=1}^C \boldsymbol{\mu}_c$. Let $W = [\mathbf{w}_1, \ldots, \mathbf{w}_C]^\top \in \mathbb{R}^{C\times d}$ denote the weight matrix of the final linear classifier. The properties of the neural collapse are as follows [17]:

1. **(NC1) Variability Collapse:** The within-class variability of last-layer features collapses to zero:

$$\Sigma_W = \mathbb{E}_{(\mathbf{x},y)\in\mathcal{D}_{id}}[(\phi_\theta(\mathbf{x}) - \boldsymbol{\mu}_y)(\phi_\theta(\mathbf{x}) - \boldsymbol{\mu}_y)^\top] \to 0 \tag{2}$$

2. **(NC2) Convergence to a Simplex ETF:** The centered class-means $\boldsymbol{\mu}_c - \boldsymbol{\mu}_G$ form a *Simplex Equiangular Tight Frame (ETF)*. That is, all class means are equidistant from the global mean and equally separated from each other:

$$\|\boldsymbol{\mu}_c - \boldsymbol{\mu}_G\| = r, \quad \langle \tilde{\boldsymbol{\mu}}_c, \tilde{\boldsymbol{\mu}}_{c'} \rangle = -\frac{1}{C-1} \quad \forall c \neq c' \tag{3}$$

where $\tilde{\boldsymbol{\mu}}_c = \frac{\boldsymbol{\mu}_c - \boldsymbol{\mu}_G}{\|\boldsymbol{\mu}_c - \boldsymbol{\mu}_G\|}$.

3. **(NC3) Self-Duality:** The classifier weights align with the class-means up to scaling:

$$\left\| \frac{W^\top}{\|W\|_F} - \frac{\dot{M}}{\|\dot{M}\|_F} \right\|_F \to 0 \tag{4}$$

where $\dot{M} = [\boldsymbol{\mu}_1 - \boldsymbol{\mu}_G, \ldots, \boldsymbol{\mu}_C - \boldsymbol{\mu}_G]$.

4. **(NC4) Simplification to Nearest Class-Center (NCC) Classification:** The classification decision reduces to comparing Euclidean distances to class means:

$$\arg\max_c \langle \mathbf{w}_c, \phi_\theta(\mathbf{x}) \rangle + b_c \to \arg\min_c \|\phi_\theta(\mathbf{x}) - \boldsymbol{\mu}_c\|^2 \tag{5}$$

5. **(NC5) Separation of ID and OOD:** The simplex ETF created due to the neural Collapse is orthogonal to the OOD inputs. Formally, if $\mathbf{x}_{ood} \sim \mathcal{P}_{ood}$, then the average feature $\mu_G^{ood} = \mathbb{E}[\phi_\theta(\mathbf{x}_{ood})]$ is orthogonal to the class mean μ_c, where $c \in [1, 2, ..., C]$ as described by Eqn. 6.

$$\frac{\langle \boldsymbol{\mu_c}, \boldsymbol{\mu_G^{ood}} \rangle}{||\boldsymbol{\mu_c}||_2 ||\boldsymbol{\mu_G^{ood}}||_2} \to 0, \forall c \tag{6}$$

While the first four properties appeared in [18], the last property related to the OOD detection was proposed by [1]. The last property suggests that ID is separable from OOD based on the measurable property of the DNN.

Definition of Energy Score. The *energy score* is a post-hoc OOD detection method proposed by [13] that leverages the connection between the output logits of a DNN and the concept of energy in statistical physics. Given a DNN classifier $\mathcal{F}_\theta(\mathbf{x})$ with logit outputs $f_1(\mathbf{x}), \ldots, f_C(\mathbf{x})$ for C classes, the energy score is defined as:

$$E(\mathbf{x}) = -T \cdot \log \sum_{c=1}^{C} \exp\left(\frac{f_c(\mathbf{x})}{T}\right) \tag{7}$$

where $T > 0$ is a temperature parameter.

This energy score arises from the energy-based formulation of the softmax probability:

$$p(y = c|\mathbf{x}) = \frac{\exp(-E_c(\mathbf{x}))}{\sum_{j=1}^{C} \exp(-E_j(\mathbf{x}))}, \tag{8}$$

where $E_c(\mathbf{x}) = -f_c(\mathbf{x})$ is the class-specific energy. The total energy $E(\mathbf{x})$ measures the compatibility between the input $\mathbf{x}$ and the model, assigning lower energy to more likely (i.e., in-distribution) samples.

In OOD detection, $S(\mathbf{x}) = -E(\mathbf{x})$ is used as a score function with the goal of setting a threshold τ to distinguish between $\mathcal{P}_{id}$ and $\mathcal{P}_{ood}$. Unlike softmax confidence, the energy score has a stronger theoretical foundation and empirically shows better separation between ID and OOD distributions.

3. Energy Score Under Neural Collapse

We begin by asking: *How does the energy score behave for ID and OOD inputs under the assumption of Neural Collapse (NC)?* To investigate this, let $\mathbf{h} = \phi(\mathbf{x})$ denote the feature from the penultimate layer of a DNN for an input $\mathbf{x}$. With a linear classifier defined by weight matrix $\mathbf{W}$ and bias $\mathbf{b}$, the resulting logits are:

$$\mathbf{z} = \mathbf{W}\mathbf{h} + \mathbf{b} \tag{9}$$

Under perfect Neural Collapse, specifically properties NC1 and NC4, the feature vector $\mathbf{h}$ collapses to the class mean $\boldsymbol{\mu}_c$, where c is the predicted class of $\mathbf{x}$. This leads to the centered class mean representation:

$$\tilde{\boldsymbol{\mu}}_c = \boldsymbol{\mu}_c - \boldsymbol{\mu}_G = \mathbf{h} - \boldsymbol{\mu}_G \tag{10}$$

where $\boldsymbol{\mu}_G$ is the global mean. Substituting into Eqn. 9, the logit for class c' becomes:

$$z_{c'} = \mathbf{w}_{c'}^T \tilde{\boldsymbol{\mu}}_c + \mathbf{w}_{c'}^T \tilde{\boldsymbol{\mu}}_G + \mathbf{b} \tag{11}$$

From NC3, the classifier weights are aligned with the centered class means. Writing $\mathbf{W} = [\mathbf{w}_1^T, \ldots, \mathbf{w}_C^T]^T$, we have:

$$\begin{aligned} \frac{\mathbf{W}^T}{\|\mathbf{W}\|_F} &= \frac{\dot{M}}{\|\dot{M}\|_F} \\ \Rightarrow \mathbf{W}^T &= \frac{\|\mathbf{W}\|_F}{\|\dot{M}\|_F} \dot{M} \\ \Rightarrow \mathbf{w}_c &= \alpha \tilde{\boldsymbol{\mu}}_c \end{aligned} \tag{12}$$

where $\dot{M}$ is the matrix of centered class means and $\alpha = \frac{\|\mathbf{W}\|_F}{\|\dot{M}\|_F}$ is a model-dependent scaling constant.

From NC1, the inner product between class means follows:

$$\boldsymbol{\mu}_c^T \boldsymbol{\mu}_{c'} = \begin{cases} r^2, & \text{if } c = c' \\ -\frac{r^2}{C-1}, & \text{if } c \neq c' \end{cases} \tag{13}$$

Substituting this into Eqn. 11, we find that the logit for class c' simplifies to:

$$z_{c'} = \begin{cases} \alpha r^2 + K, & \text{if } c = c' \\ -\frac{\alpha r^2}{C-1} + K, & \text{otherwise} \end{cases} \tag{14}$$

where $K = \alpha \boldsymbol{\mu}_{c'}^T \boldsymbol{\mu}_G + \mathbf{b}$ is a constant shared across all classes.

This yields the energy score for ID inputs:

$$\begin{aligned} E_{\text{ID}}(\mathbf{x}) &= -\log \left[\exp(\alpha r^2) + \sum_{i=1}^{C-1} \exp\left(-\frac{\alpha r^2}{C-1}\right) \right] - K \\ &= -\log \left[\exp(\alpha r^2) + (C-1) \exp\left(-\frac{\alpha r^2}{C-1}\right) \right] - K \end{aligned} \tag{15}$$

This expression is constant for all ID inputs under perfect NC. To compute the energy score for OOD inputs, we consider the fifth property of Neural Collapse, which implies orthogonality between any class mean $\boldsymbol{\mu}_c$ and the OOD feature vector $\boldsymbol{\mu}^{\text{ood}}$, i.e., $\boldsymbol{\mu}_c^T \boldsymbol{\mu}^{\text{ood}} = 0$. Since $\mathbf{w}_c \propto \boldsymbol{\mu}_c$, it follows that $\mathbf{w}_c^T \boldsymbol{\mu}^{\text{ood}} = 0$ for all c. Hence, the logit for every class becomes:

$$z_{c'} = K \tag{16}$$

which gives the following energy score for OOD inputs:

$$E_{\text{OOD}}(\mathbf{x}) = -\log\left[\sum_{i=1}^{C} \exp(K)\right] = -K + \log(C) \tag{17}$$

Thus, the energy gap between ID and OOD inputs is:

$$\begin{aligned}\Delta E &= E_{\text{IID}} - E_{\text{ID}} \\ &= \log\left[\frac{\exp(\alpha r^2)}{C} + \left(1 - \frac{1}{C}\right)\exp\left(-\frac{\alpha r^2}{C-1}\right)\right]\end{aligned} \tag{18}$$

If αr^2 is sufficiently large, the second term becomes negligible, and the gap simplifies to:

$$\Delta E \approx \alpha r^2 - \log(C) \tag{19}$$

This implies that the separation between ID and OOD energy scores increases with the norm of the class means r, which also corresponds to the feature norm. Therefore, feature scaling can significantly improve energy-based separation under the Neural Collapse regime.

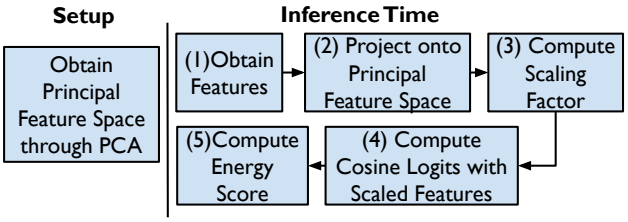

Figure 2. Steps for OOD detection with ENCORE.

3.1. ENCORE for OOD Detection

Motivated by the observation that, under the Neural Collapse (NC) regime, scaling feature vectors improves separation between ID and OOD inputs, we propose ENCORE, a principled approach for energy-based OOD detection. Since practical DNNs are rarely trained to full convergence, the assumptions of perfect Neural Collapse may not hold. To address this, ENCORE aims to modify the model's behavior at inference time to more closely emulate the structure imposed by NC.

Recall that the third NC property (NC3) asserts that the classifier weights are aligned with the class means. When combined with the first NC property (NC1), which states that all features from a given class collapse to their respective class mean, this suggests that the weight vector corresponding to the predicted class should be linearly aligned with the feature vector. Therefore, the cosine similarity between the feature and class weight vectors emerges as a natural surrogate for the logit.

Beyond alignment, this formulation offers a key practical benefit: NC1 also implies that feature vectors should be of equal norm, a condition that is often violated in practice. By using cosine similarity, we remove norm dependence from the logits, isolating the angular component that better reflects class identity. The benefit of normalization is also emphasized in works like [11, 20]. Furthermore, introducing an explicit control over the feature norm via a scaling parameter κ allows us to modulate the separation between ID and OOD inputs. The resulting logit (which will henceforth be referred as cosine logits) for class c is defined as:

$$z_c = \kappa \cos\theta_c = \kappa \frac{\mathbf{w}_c^T \mathbf{h}(\mathbf{x})}{\|\mathbf{w}_c\|_2 \|\mathbf{h}(\mathbf{x})\|_2} \tag{20}$$

Here, θ_c is the angle between the feature and weight vector for class c. This leads to the following energy score:

$$E(\mathbf{x}) = -\log\left[\sum_{i=1}^{C} \exp(\kappa \cos\theta_i)\right] \tag{21}$$

While this construction does not explicitly enforce the ETF simplex structure posited by NC2, it does yield an equinorm angular configuration, which empirically enhances robustness. In proposition 1, we formally show, under assumption of deviation from the assumptions of neural collapse, how the energy gap between ID and OOD gets affected by the scaling factor, and assumed deviations.

Proposition 1 *Let cosine logits be of the form $z_c = \kappa\cos(\theta_c)$, as defined in Eqn. 20. Assume: i) the weights $\{w_c\}_{c=1}^{C}$ and class means $\boldsymbol{\mu}_c$ form a unit-norm Equiangular Tight Frame (ETF), with pairwise inner product $\rho < 1$ ii) an ID input has angle $\theta \approx 0$ with its correct class weight w_y, so $z_y = \kappa\cos(\theta)$, and for all $c \neq y$, $z_c = \kappa\rho$ iii) an OOD input has approximately equal angle $\Phi \approx 90°$ with all class weights, so $z_c = \kappa\cos(\Phi) \approx \kappa\epsilon$. Then the energy gap between ID and OOD satisfies:*

$$\begin{aligned}\Delta E &= E_{\text{ood}}(\epsilon) - E_{\text{id}}(\theta) \\ &\leq -\log C + \kappa\left(1 - \frac{\theta^2}{2} - \epsilon\right)\end{aligned}$$

We provide the proof in the supplementary section 1. This shows that increasing the deviation of the features from the class means (represented by θ) or decreasing the angle between OOD features and class means (represented by Φ and ϵ) reduces the separation between ID and OOD leading to increased false negatives.

To further refine the energy score, we leverage NC5, which suggests that OOD features are ideally orthogonal to the span of ID class means. This property has been utilized in prior work [1, 21] to derive orthogonal projections for OOD detection. In ENCORE, we incorporate this idea by making the feature scaling parameter *sample-adaptive*. Specifically, we scale the cosine logits by the ratio of the projection of the feature vector onto the ID subspace to its total norm, multiplied by a fixed constant λ (which we will be referred as scaling constant). Let $\|\mathbf{x}^p\|$ denote the norm of the projection of $\mathbf{x}$ onto the ID subspace. Then the scaling factor is given by $\kappa = \exp\left(\lambda \frac{\|\mathbf{x}^p\|}{\|\mathbf{x}\|}\right)$ and the final energy score used in ENCORE becomes:

$$\begin{aligned} E(\mathbf{x}) &= -\log\left[\sum_{i=1}^{C} z_i\right] \\ &= -\log\left[\sum_{i=1}^{C} \exp\left(\exp\left(\lambda \frac{\|\mathbf{x}^p\|}{\|\mathbf{x}\|}\right)\cos\theta_i\right)\right] \end{aligned} \tag{22}$$

This adaptive rescaling enhances the separation between ID and OOD inputs by attenuating the influence of features that are weakly aligned with the ID feature space. A potential concern with this approach is numerical overflow under FP16 precision, particularly due to the exponential scaling. However, from Eqn.20, we observe that the magnitude of the scaled logit is bounded as $|z_c| \leq \kappa$. This ensures numerical stability as long as κ and in turn the scaling constant λ remains within a safe range. In practice, we find that setting λ between 0 and 11 does not lead to overflow during logit computation. Moreover, as demonstrated in Fig.4 and discussed in Sec. 5, the optimal value of the scaling constant λ converges rapidly due to the exponential function and typically remains around 3 which is well below the critical threshold. Therefore, the proposed scaling mechanism is both effective and numerically stable in FP16 environments.

To summarize the steps for OOD detection with ENCORE (as illustrated in Figure 2): during setup phase, we extract penultimate layer features and compute the principal feature subspace. During inference: 1) we extract the penultimate layer feature 2) project the feature onto principal feature space 3) compute scaling factor κ 4) compute cosine logits according to Equation 20 5) compute OOD score with Equation 22.

4. Experimental Results

In this section, we evaluate the performance of our method ENCORE across OOD benchmarks against diverse baselines. In line with existing literature, we compute the False Positive Rate at 95% true positive rate (FPR95) and Area Under Receiver Operating Curve (AUROC). A lower FPR95 and a higher AUROC indicates better performance. Unless stated otherwise, FPR means the same as FPR95.

OOD Benchmarks. We adopt the widely known OpenOOD [24, 27] benchmark. We run for both the CIFAR and Imagenet benchmarks of OpenOOD. For each benchmark, the OOD datasets are split into two groups - *near-OOD* and *far-OOD*. We report for both the individual datasets and the aggregate performance on near-OOD and far-OOD.

DNN Architectures. We test CIFAR on Resnet 18 [7] and use RepVGG [2] as an alternate architecture. To test performance of ENCORE for both convolutional and transformer architectures, we run experiments for the Imagenet benchmark on the ConvNext [15] and use Vision Transformer [4] as an alternate architecture.

4.1. Evaluation on CIFAR-10

Table 1 compares ENCORE with the baselines for Resnet 18 architecture. We can observe that ENCORE achieves state-of-the-art performance in terms of FPR95 and AUROC. To better understand the position of ENCORE with respect to the baselines, we highlight some specific groups.

ENCORE vs MSP/Energy/GEN: All the three baselines operate in the logit space. MSP relies on the maximum softmax probability, discarding information about rest of the classes. Energy, in contrast, utilizes logits from all classes to better separate ID and OOD inputs. GEN takes an intermediate approach, demonstrating that using the top *M* softmax probabilities improves OOD detection.

Unlike these methods, which directly uses functions of the output logits for OOD detection, ENCORE enhances the separability between ID and OOD by scaling in the feature space. This is evident from the substantial improvement of ENCORE over MSP by 17.34% in FPR95 on near OOD datasets. Similarly, ENCORE outperforms Energy and GEN by 31.91% and 21.62%, respectively. For far OOD datasets, ENCORE achieves average improvements of 10.11%, 19.3%, and 12.48% over MSP, Energy, and GEN. This proves usefulness of ENCORE compared to these approaches on CIFAR10 benchmark.

ENCORE vs KNN/FDBD For benchmarking KNN, we follow the hyperparameter setup from [20], using $k = 50$ nearest neighbors across the entire training dataset. KNN relies on computing the average distance to these neighbors, requiring storage of the entire training set's features. In contrast, ENCORE only uses the train set for estimating the feature space through Principal Component Analysis (PCA), drastically reducing memory overhead while simplifying the detection process. For a fixed number of classes, required memory of ENCORE is constant while for KNN, it increases with the size of the dataset. Despite being significantly more lightweight, ENCORE outperforms KNN, albeit slightly, in terms of AUROC. For KNN, this improvement is 1.43% for far-OOD benchmark and 0.5%

Table 1. Comparison of ENCORE with baseline approaches on the CIFAR10 benchmark using ResNet-18, showing results for Near-OOD and Far-OOD datasets.

	Near OOD		Far OOD	
	FPR ↓	AUROC ↑	FPR ↓	AUROC ↑
MSP	53.61	87.69	31.22	91.03
OpenMax	47.19	87.20	29.32	89.55
ODIN	84.71	80.37	61.02	87.24
MDS	46.04	86.79	30.26	90.21
RMDS	42.26	89.57	24.31	92.46
GRAM	93.88	52.43	69.34	69.70
Energy	68.18	86.95	40.41	91.80
GradNorm	95.26	53.93	89.35	58.65
React	71.07	86.51	42.08	91.09
ViM	47.69	88.55	25.68	93.17
KNN	39.42	88.73	23.73	93.13
DICE	80.52	77.68	53.78	85.44
ASH	88.98	74.20	76.36	78.46
GEN	57.89	87.80	33.59	91.57
FDBD	36.62	90.45	23.49	93.20
CoRP	38.60	90.46	22.48	94.12
ENCORE (Ours)	**36.27**	**91.23**	**21.11**	**94.56**

for near-OOD benchmark. We observe a similar case for the FDBD which uses the train set to compute the average feature. ENCORE outperforms FDBD by 0.77% on near-OOD and 1.36% on far-OOD benchmark.

ENCORE vs ViM The similarity between ViM and ENCORE is that both employ the information related to principal feature-space of the training inputs. While ViM uses the energy in the null-space, ENCORE focuses on the principal feature space. Apart from that, ViM uses the energy in the null-space of the features as an extra logit. ENCORE uses the percentage of energy in the feature space for adaptively setting the scaling factor. As can be seen from Table 1, this improves the OOD detection significantly for CIFAR10 benchmark. For near-OOD, this improves by 11.42% in terms of FPR. For far-OOD, this improves by 4.57%.

Table 2. Comparison of ENCORE with baseline approaches on the ImageNet benchmark using ConvNeXt-Small, showing results for Near-OOD and Far-OOD datasets.

	Near OOD		Far OOD	
	FPR ↓	AUROC ↑	FPR ↓	AUROC ↑
MSP	75.74	77.58	69.98	83.12
React	75.79	72.61	71.11	77.55
ViM	70.13	**78.66**	28.36	90.71
KNN	67.60	77.90	**28.27**	**92.23**
DICE	94.41	59.96	84.04	71.91
ASH	98.70	42.29	97.24	49.43
GEN	67.08	77.33	45.95	90.08
CoRP	67.51	72.13	29.82	89.89
ENCORE (Ours)	**66.01**	77.70	28.53	90.72

4.2. Evaluation on Imagenet Benchmark

Table 2 presents a comprehensive comparison of ENCORE against state-of-the-art OOD detection baselines on the ImageNet benchmark using the ConvNeXt backbone. Across a wide range of OOD datasets - including both near and far shifts - ENCORE demonstrates consistent and superior performance. On the *Near OOD* splits, ENCORE achieves the lowest FPR (66.01%), outperforming all baselines, including ViM and KNN, while maintaining competitive AUROC score (77.70%). This suggests that ENCORE is effective in detecting subtle distribution shifts where OOD inputs resemble ID samples. For the *Far OOD* sets, on average, ENCORE again achieves competitive FPR (28.53% compared to the best 28.27%) and strong AUROC scores (90.72% aas compared to the best 92.23%), showing its robustness even when the semantic gap between ID and OOD data is large. From the detailed resutls provided in the supplementary Sec. 5, we observe that **ENCORE shows strong performance across different benchmarks**, outperforming baselines that rely on activation suppression (ASH), feature thresholding (React), and logit-based uncertainty (GEN). The neural-collapse inspired design of ENCORE allows it to achieve this competitive or superior performance across benchmarks.

4.3. Comparison of Latency

We evaluate the inference latency of ENCORE and compare it with two of the closest performing OOD detection methods- ViM and KNN. All measurements are conducted on the RepVGG-a2 architecture using an NVIDIA A100 GPU with a batch size of 1024. As shown in Table 3, ENCORE achieves the lowest latency at 135,ms, significantly outperforming both ViM (155,ms) and KNN (354,ms). This demonstrates that ENCORE not only provides competitive detection performance but also offers superior computational efficiency, making it well-suited for high-throughput deployment scenarios.

Table 3. Inference latency (ms) for different OOD detection methods on RepVGG-a2 with batch size 1024.

Method	**Latency (ms)**
ENCORE	135
ViM	155
KNN	354

4.4. Ablation Study

4.4.1 Number of principal components

We perform ablation study to understand the effect of the dimension of the principal feature-space. As can be observed from Figure 3, increasing the dimension (number of components of principal component analysis) drops the FPR. With above 300 dimensions, the FPR stabilizes. In general, we observe that for stable performance, the number of dimension should be chosen to achieve at least 95% explained

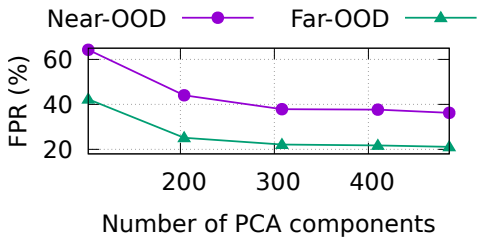

Figure 3. Variation of FPR with varying number of dimensions for principal feature-space. The results are for ResNet18 and CIFAR10 benchmark.

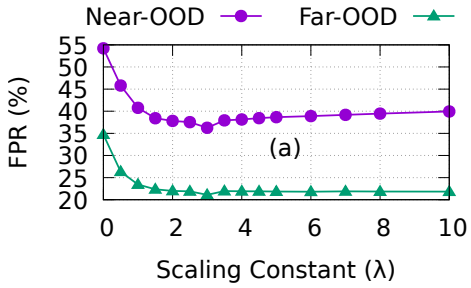

Figure 4. Variation of FPR with varying the scaling constant (λ)

variance of the ID features. In our experiments, we keep the explained variance 99%.

4.4.2 Varying the scaling factor

The key observation ENCORE is that uniformly scaling up the features can achieve good separation between ID and OOD. We run ablation study to understand the effect of the scaling constant (λ) which controls the scaling factor. We vary the scaling constant between 0 and 10. We observe that with scaling factor of zero (0) - no scaling - the FPR is highest. Increasing the scaling constant quickly drops the FPR and it saturates at around value of 3. In all our experiments, we use a scaling constant of 3.

4.4.3 Ablation Study for Different Components of ENCORE

To better approximate the equinorm structure associated with neural collapse, ENCORE incorporates both feature norm scaling and cosine similarity between the scaled feature vectors and class weight vectors. Table 4 presents an ablation study of these components on the ImageNet benchmark using the ViT-Huge architecture. The results show that using cosine similarity alone yields the weakest performance in both near- and far-OOD scenarios. Feature norm scaling alone leads to a noticeable improvement, but the best performance is achieved when both components are combined. This indicates that the joint use of cosine similarity and norm scaling more effectively induces an ETF-like geometry in the feature space, thereby enhancing the separation between ID and OOD inputs.

Table 4. Ablation study for different components of ENCORE on the ImageNet benchmark.

	Near OOD		Far OOD	
	FPR ↓	AUROC ↑	FPR ↓	AUROC ↑
Only Norm Scaling	59.68	85.15	37.76	90.37
Only Cosine Similarity	79.27	73.88	60.85	78.32
Cosine Similarity+Norm Scaling	**43.55**	**89.47**	**14.62**	**96.89**

4.5. Full-spectrum OOD Detection

We evaluate ENCORE under the *Full-Spectrum Out-of-Distribution (FS-OOD)* detection framework [26], which distinguishes semantic OOD samples from those caused by mere covariate shifts. In this setup, covariate-shifted inputs are considered ID, making FS-OOD a more challenging and realistic benchmark. Table 5 highlights that ENCORE sets a new state-of-the-art on the Imagenet benchmark across both near-OOD and far-OOD categories. Specifically, ENCORE achieves the best AUROC in 3 out of 5 OOD datasets and the lowest FPR in 4 out of 5, indicating its strong capability to separate semantic anomalies while being tolerant to covariate shifts. In the near-OOD regime, ENCORE significantly outperforms all baselines, achieving the lowest FPR (79.61%) and highest AUROC (59.63%), showing its precision in not misclassifying covariate-shifted data. On the far-OOD datasets such as Openimage-O and Textures, ENCORE also secures top AUROC scores (77.10% and 73.54%, respectively), outperforming established methods like ViM and GEN. Notably, while methods like ASH and DICE occasionally exhibit lower FPR in isolated cases, they suffer from drastically poor AUROC, often below 50%, indicating poor discrimination. In contrast, ENCORE achieves a consistent trade-off, maintaining high AUROC scores without excessively high FPRs, thereby excelling in the FS-OOD setting. Overall, while no single baseline performs consistently well across all metrics, ENCORE emerges as the most reliable method across the full OOD spectrum.

4.6. Results on Alternate Architectures

As shown in Tab. 7, ENCORE performs very strongly even for vision transformer architecture. It consistently outperforms the SOTA approaches like [14], [6], [5] on both near- and far-OOD settings by upto 1.85% in terms of FPR. This serves to establish that ENCORE works for both convolutional and transformer architectures making it model-agnostic and dependent on DNN's properties only. Similarly, for the CIFAR benchmark, we observe in Tab. 6 that ENCORE outperforms the SOTA approaches for RepVGG-a2 architecture.

4.7. Contrast with Similar Approaches

Contrast with Feature Scaling Approaches Several OOD detection methods aim to reshape features based

Table 5. Evaluation of ENCORE in the *Full-spectrum OOD Detection* setup on the ImageNet benchmark using ConvNeXt-Small. Results are grouped by Near-OOD and Far-OOD datasets as per the FS-OOD protocol.

	SSB-Hard		NINCO		Near OOD		iNaturalist		Textures		OpenImage-O		Far OOD	
	FPR ↓	AUROC ↑	FPR ↓	AUROC ↑	FPR ↓	AUROC ↑	FPR ↓	AUROC ↑	FPR ↓	AUROC ↑	FPR ↓	AUROC ↑	FPR ↓	AUROC ↑
MSP	89.22	55.17	83.51	62.71	86.37	58.94	62.47	74.91	96.48	59.78	86.30	65.80	81.75	66.83
React	91.11	47.68	83.94	55.40	87.52	51.54	76.10	55.89	93.31	58.49	84.36	57.29	84.59	57.23
ViM	91.69	**51.23**	79.46	63.72	85.57	57.48	**47.62**	78.28	63.93	**74.76**	**56.21**	69.70	55.92	76.71
KNN	91.10	49.69	76.09	61.90	82.60	55.80	50.88	77.40	64.39	70.17	55.66	75.14	**56.97**	**74.24**
DICE	95.11	46.54	96.85	47.36	95.98	46.95	91.04	51.00	85.51	65.82	89.69	58.28	88.74	58.36
ASH	97.30	40.02	96.44	43.62	96.87	41.82	96.04	36.85	90.34	62.79	97.38	45.19	94.59	**77.71**
GEN	90.97	48.00	75.55	58.84	83.26	53.42	**47.70**	**80.92**	91.22	65.30	65.99	71.18	68.31	72.47
ENCORE (Ours)	**85.33**	**54.75**	**73.89**	**64.50**	**79.61**	**59.63**	52.84	66.27	**60.80**	73.54	**54.17**	**77.10**	55.94	72.30

Table 6. Comparison of ENCORE with baseline approaches on the CIFAR10 benchmark using RepVGG-A2, showing results for Near-OOD and Far-OOD datasets.

	Near OOD		Far OOD	
	FPR ↓	AUROC ↑	FPR ↓	AUROC ↑
Energy	72.07	86.94	49.18	90.89
GradNorm	98.33	40.24	89.36	58.47
React	57.97	80.16	58.34	77.57
ViM	38.47	89.57	29.01	91.56
KNN	36.20	90.70	28.46	91.95
DICE	79.84	82.63	54.70	89.21
ASH	97.03	60.06	88.74	69.09
GEN	70.96	87.05	48.71	90.77
FDBD	**34.24**	90.52	29.55	91.89
CoRP	36.35	89.50	31.14	90.87
ENCORE (Ours)	34.72	**90.77**	**28.35**	**92.89**

Table 7. Comparison of ENCORE with baseline approaches on the ImageNet benchmark using ViT-Huge, showing results for Near-OOD and Far-OOD datasets.

	Near OOD		Far OOD	
	FPR ↓	AUROC ↑	FPR ↓	AUROC ↑
MSP	56.47	84.15	31.51	93.27
React	44.86	88.99	15.11	96.33
ViM	45.41	88.77	21.32	93.74
KNN	61.54	80.64	26.03	94.16
DICE	67.01	77.01	39.56	87.95
ASH	84.99	64.18	72.99	76.43
GEN	44.14	88.86	14.69	96.66
CoRP	45.40	87.78	15.26	96.15
ENCORE (Ours)	**43.55**	**89.47**	**14.62**	**96.89**

on training or instance-wise statistics to improve ID-OOD separability. [19] introduced feature thresholding using ID training statistics, while [3] enhanced high-magnitude features and suppressed low-magnitude ones. However, [23] showed that such suppression can harm OOD performance, proposing selective scaling instead. In contrast, ENCORE avoids thresholding or selective scaling and instead applies uniform feature scaling, which proves effective under neural collapse assumptions. This is further improved by using cosine similarity with class weights as a proxy for logits. Performance differences among these methods are shown in Table 8.

Contrast with NECO [1] first utilized the concept of neural collapse for OOD detection. The fifth property of neural collapse is suggested by [1]. The core idea for OOD detection by [1] was to use the ratio (scaled by the maximum logit) between the norm of the features projected onto the principal feature space and the norm of the features themselves. ENCORE takes a completely different approach. ENCORE still uses this ratio as according to the fifth property of neural collapse, this can discriminate between the ID and OOD inputs. But this is used to adaptively scale the features. The core ideas of ENCORE is: i) we can scale the features to better separate the ID and OOD inputs based on the properties of neural collapse ii) using cosine-similarity between the features and weight vectors is better proxy for the logits for OOD detection. The difference in approach with [1] also gets clear from the difference in performance as highlighted in Table 8.

Table 8. Comparison of ENCORE with similar approaches on the ImageNet benchmark using ViT-Huge-14, showing results for Near-OOD and Far-OOD datasets.

	Near OOD		Far OOD	
	FPR ↓	AUROC ↑	FPR ↓	AUROC ↑
SCALE	58.24	77.52	26.53	88.51
NECO	55.55	79.64	24.25	89.96
ENCORE (Ours)	**43.55**	**89.47**	**14.62**	**96.89**

5. Conclusion

In this work, We revisited energy-based OOD detection through the lens of neural collapse and showed that feature scaling can improve ID-OOD separation. Building on this insight, we proposed ENCORE, a post-hoc detection method that combines cosine similarity and adaptive feature scaling to approximate the geometry of neural collapse. ENCORE achieves competitive or superior performance across multiple benchmarks and architectures, all without requiring model retraining or architectural modifications. While ENCORE focuses on properties induced by neural collapse, future work may explore combining it with other structural traits of deep networks or adapting its principles to non-classification tasks, broadening its applicability.

Acknowledgements

This work has been supported by the National Science Foundation under grants CNS-2312875 and OAC-2530896; by the Air Force Office of Scientific Research under grant FA9550-23-1-0261; by the Office of Naval Research under grant N00014-23-1-2221; and by the Defense Advanced Research Projects Agency under Cooperative Agreement D25AC00374-00.

References

[1] Mouïn Ben Ammar, Nacim Belkhir, Sebastian Popescu, Antoine Manzanera, and Gianni Franchi. Neco: Neural collapse based out-of-distribution detection, 2023. 3, 5, 8

[2] Xiaohan Ding, Xiangyu Zhang, Ningning Ma, Jungong Han, Guiguang Ding, and Jian Sun. Repvgg: Making vgg-style convnets great again. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 13733–13742, 2021. 5

[3] Andrija Djurisic, Nebojsa Bozanic, Arjun Ashok, and Rosanne Liu. Extremely simple activation shaping for out-of-distribution detection. In *The Eleventh International Conference on Learning Representations*, 2022. 1, 8

[4] 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. 5

[5] Kun Fang, Qinghua Tao, Mingzhen He, Kexin Lv, Runze Yang, Haibo Hu, Xiaolin Huang, Jie Yang, and Longbin Cao. Kernel pca for out-of-distribution detection: Non-linear kernel selections and approximations, 2025. 7

[6] Kun Fang, Qinghua Tao, Kexin Lv, Mingzhen He, Xiaolin Huang, and JIE YANG. Kernel PCA for out-of-distribution detection. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024. 7

[7] Kaiming He, X. Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 770–778, 2015. 5

[8] Dan Hendrycks, Steven Basart, Mantas Mazeika, Andy Zou, Joseph Kwon, Mohammadreza Mostajabi, Jacob Steinhardt, and Dawn Song. Scaling out-of-distribution detection for real-world settings. In *International Conference on Machine Learning*, pages 8759–8773. PMLR, 2022. 1

[9] Dan Hendrycks and Kevin Gimpel. A baseline for detecting misclassified and out-of-distribution examples in neural networks. In *International Conference on Learning Representations*, 2016. 1

[10] 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*, 2021. 1

[11] Yuting Li, Yingyi Chen, Xuanlong Yu, Dexiong Chen, and Xi Shen. Sure: Survey recipes for building reliable and robust deep networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 17500–17510, June 2024. 4

[12] Shiyu Liang, Yixuan Li, and R. Srikant. Enhancing the reliability of out-of-distribution image detection in neural networks. In *International Conference on Learning Representations*, 2018. 1

[13] 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, 3

[14] Xixi Liu, Yaroslava Lochman, and Christopher Zach. Gen: Pushing the limits of softmax-based out-of-distribution detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 23946–23955, 2023. 7

[15] Zhuang Liu, Hanzi Mao, Chao-Yuan Wu, Christoph Feichtenhofer, Trevor Darrell, and Saining Xie. A convnet for the 2020s. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 11976–11986, 2022. 5

[16] Yifei Ming, Yiyou Sun, Ousmane Dia, and Yixuan Li. How to Exploit Hyperspherical Embeddings for Out-of-Distribution Detection? In *The Eleventh International Conference on Learning Representations*, 2023. 1

[17] Vardan Papyan, XY Han, and David L Donoho. Prevalence of neural collapse during the terminal phase of deep learning training. *Proceedings of the National Academy of Sciences*, 117(40):24652–24663, 2020. 1, 2

[18] Vardan Papyan, Xiuyuan Li Han, and David Donoho. Prevalence of neural collapse during the terminal phase of deep learning training. In *International Conference on Learning Representations (ICLR)*, 2021. 3

[19] 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. 1, 8

[20] Yiyou Sun, Yifei Ming, Xiaojin Zhu, and Yixuan Li. Out-of-distribution Detection with Deep Nearest Neighbors. *ICML*, 2022. 4, 5

[21] 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. 1, 5

[22] Hongxin Wei, Renchunzi Xie, Hao Cheng, Lei Feng, Bo An, and Yixuan Li. Mitigating neural network overconfidence with logit normalization. In *International Conference on Machine Learning*, pages 23631–23644. PMLR, 2022. 1

[23] Kai Xu, Rongyu Chen, Gianni Franchi, and Angela Yao. Scaling for training time and post-hoc out-of-distribution detection enhancement. In *The Twelfth International Conference on Learning Representations*, 2024. 8

[24] Jingkang Yang, Pengyun Wang, Dejian Zou, Zitang Zhou, Kunyuan Ding, Wenxuan Peng, Haoqi Wang, Guangyao Chen, Bo Li, Yiyou Sun, Xuefeng Du, Kaiyang Zhou, Wayne Zhang, Dan Hendrycks, Yixuan Li, and Ziwei Liu. Openood: Benchmarking generalized out-of-distribution detection. 2022. 5

[25] Jingkang Yang, Kaiyang Zhou, Yixuan Li, and Ziwei Liu. Generalized out-of-distribution detection: A survey. *arXiv e-prints*, pages arXiv–2110, 2021. 1

[26] Jingkang Yang, Kaiyang Zhou, and Ziwei Liu. Full-spectrum out-of-distribution detection. *arXiv preprint arXiv:2204.05306*, 2022. 7

[27] Jingyang Zhang, Jingkang Yang, Pengyun Wang, Haoqi Wang, Yueqian Lin, Haoran Zhang, Yiyou Sun, Xuefeng Du, Yixuan Li, Ziwei Liu, Yiran Chen, and Hai Li. Openood v1.5: Enhanced benchmark for out-of-distribution detection. *arXiv preprint arXiv:2306.09301*, 2023. 5