GROOD: Gradient-Aware Out-of-Distribution Detection

Anonymous authors
Paper under double-blind review

Abstract

Out-of-distribution (OOD) detection is crucial for ensuring the reliability of deep learning models in real-world applications. Existing methods typically focus on feature representations or output-space analysis, often assuming a distribution over these spaces or leveraging gradient norms with respect to model parameters. However, these approaches struggle to distinguish near-OOD samples and often require extensive hyper-parameter tuning, limiting their practicality. In this work, we propose GRadient-aware Out-Of-Distribution detection (GROOD), a method that derives an OOD prototype from synthetic samples and computes class prototypes directly from In-distribution (ID) training data. By analyzing the gradients of a nearest-class-prototype loss function concerning an artificial OOD prototype, our approach achieves a clear separation between in-distribution and OOD samples. Experimental evaluations demonstrate that gradients computed from the OOD prototype enhance the distinction between ID and OOD data, surpassing established baselines in robustness, particularly on ImageNet-1k. These findings highlight the potential of gradient-based methods and prototype-driven approaches in advancing OOD detection within deep neural networks.

1 Introduction

Deep neural networks (DNNs) have demonstrated exceptional performance across domains such as computer vision, natural language processing, and robotics (Goodfellow et al., 2016; LeCun et al., 2015). Their success largely relies on the assumption that training and test data follow an independent and identically distributed (iid) pattern (Krizhevsky et al., 2012; Simonyan & Zisserman, 2015). However, this assumption often fails in real-world scenarios, where DNNs encounter out-of-distribution (OOD) inputs that deviate significantly from the training distribution (Hendrycks & Gimpel, 2016). As a result, models that perform well on in-distribution (ID) data frequently produce overly confident yet incorrect predictions on OOD samples, posing significant risks to safety-critical applications such as healthcare and autonomous driving (Litjens et al., 2017; Bojarski et al., 2016). In such scenarios, it becomes imperative for the model to exhibit self-awareness about its own limitations (Gal & Ghahramani, 2015). Conventional approaches that focus solely on minimizing the training loss are often ill-equipped to cope with OOD samples, thereby jeopardizing the safe and reliable deployment of deep learning systems (Duchi & Namkoong, 2018; Arjovsky et al., 2019; Shen et al., 2019; Liu et al., 2021a).

Consequently, several active lines of research work toward equipping DNNs with the capability to effectively detect unknown or OOD samples. Among these, post-hoc OOD detection methods stand out as the most convenient, as they utilize the representations of a pre-trained DNNs, require no additional training, and can be applied to any neural network. Experimental studies using large benchmarks have underscored the effectiveness of post-hoc OOD detection methods, which has led to the development of several specialized OOD libraries (Yang et al., 2022; Zhang et al., 2023a; Kirchheim et al., 2022; Novello et al., 2023).

Existing OOD detection methods typically focus on either feature-space-based measures, which assess the distance of inputs to learned feature representations (Sun et al., 2022; Lee et al., 2018), or gradient information (Lee et al., 2022; Sun et al., 2021a; Lee & AlRegib, 2020; Chen et al., 2023), which analyze the model's gradient space. However, these methods often struggle in scenarios where OOD samples lie near class boundaries or exhibit characteristics similar to hard in-distribution ID examples. Furthermore, current approaches rarely exploit the inherent geometry of learned feature manifolds, limiting their robustness and generalization.

To address these challenges, we propose a novel post-hoc method called GRadient-aware Out-Of-Distribution detection (GROOD) that relies on feature and gradient spaces for improved OOD discrimination.

~~Our approach is built upon key insights from recent advancements in understanding neural network behavior, particularly the concept of neural collapse (NC). This property suggests that prototype-based methods, like GROOD, can be highly effective when in-distribution data exhibits well-defined cluster structure. We leverage these insights to develop a more effective way of distinguishing between ID and OOD samples in the feature space of pre-trained networks. The core idea behind GROOD is to examine how sensitive the distances between sample representations and class prototypes are to changes in an artificially introduced OOD prototype. By analyzing this sensitivity through gradients, we can effectively differentiate between ID and OOD samples without the need for additional training.~~

GROOD is specifically designed to address three persistent challenges in post-hoc and gradient-based OOD detection. First, it improves the detection of near-OOD samples those that are semantically close to the in-distribution by leveraging gradient vectors with respect to an artificial OOD prototype, which provide a more discriminative signal than feature or logit-based scores. Second, GROOD generalizes robustly across diverse architectures, including ResNets and Vision Transformers, where many existing methods suffer performance degradation. Third, it significantly reduces hyper-parameter sensitivity and exhibits stable AUROC performance across training epochs, alleviating the common issue of checkpoint instability. These properties make GROOD a practical and reliable post-hoc framework for real-world deployment.

Our method is inspired by two key observations:

1. The NC property (Papyan et al., 2020) suggests that at the end of neural network training, within-class variability tends to zero for sample representation in the feature space. This observation motivates the use of nearest class prototype (NCP) classification, which relies on the distance to class prototypes, defined as the means of the samples of each class in the feature space. To further enhance the discriminative power of this approach for out-of-distribution detection, we construct logits by incorporating distances to an additional OOD prototype, alongside the distances to the class prototypes.
2. We also observe that OOD samples tend to exhibit a more dispersed distribution in the feature space compared to ID samples. Capitalizing on this characteristic, we introduce an artificial OOD prototype, strategically positioned to be distinct from the ID class prototypes. By then examining how sample representations respond to this OOD prototype, specifically by analyzing gradients of the NCP loss with respect to it, we can gain a more nuanced understanding of the differences between ID and OOD samples, leading to more effective discrimination in the gradient space.

Our approach differs from traditional post-hoc methods by computing gradients with respect to an artificial OOD prototype rather than the network's parameters. The magnitude of these gradients serves as a key indicator: for ID data, the OOD prototype has a relatively small influence on the confidence of the prediction in the feature space, reflecting stable classification. In contrast, for OOD data, the OOD prototype has a more substantial influence on the confidence of the prediction; meaning that a smaller shift in the OOD prototype's representation is sufficient to cause a larger change in the classification confidence.

We conduct an extensive empirical study following the recent methodology introduced in the OpenOOD Benchmark (Zhang et al., 2023a), but we also evaluate our method on other recent architectures.

Our key results and contributions are summarized as follows.

- We propose GROOD, a gradient-aware OOD detection framework that integrates neural collapse geometry, gradient-space analysis, and synthetic OOD generation for robust OOD discrimination.
- We demonstrate, via an oracle experiment, that an idealized OOD prototype significantly improves OOD detection.

- We introduce a novel mixup-based approach for generating synthetic OOD data, enhancing ID/OOD decision boundaries and reducing the need for additional auxiliary OOD data.
- We conduct extensive empirical evaluations and ablations, demonstrating GROOD's effectiveness and providing new insights into the interplay of feature and gradient spaces for OOD detection.

While Neural Collapse is often viewed as a limitation for OOD detection since tightly clustered ID features can lead to overconfident misclassifications GROOD turns this behavior into a strength. By exploiting the geometric regularity of class prototypes formed under NC, GROOD identifies OOD samples not through raw confidence scores but through their abnormal gradient sensitivity to a fixed synthetic OOD prototype. This allows GROOD to leverage the structure imposed by NC while sidestepping the typical overconfidence failure mode.

2 Related Work

Neural Network Properties Prior research has emphasized the significance of linear interpolation within manifold spaces, with applications ranging from word embeddings (Mikolov et al., 2013) to machine translation (Hassan et al., 2017). Extending these concepts, Verma et al. (2019) proposed manifold mixup, a method that smooths decision boundaries and reduces overconfidence near ID data. Our work primarily leverages the NC property (Papyan et al., 2020; Kothapalli et al., 2022) and prototype/centroid-based classification. Specifically, we exploit the sensitivity of the OOD prototype to enhance the distinction between ID and OOD samples. To construct this OOD prototype, we employ manifold mixup to generate a synthetic OOD dataset, enabling a more robust and structured detection framework.

History of OOD Detection The study of handling OOD samples has a long history, dating back to early works on classification with rejection (Chow, 1970; Fumera & Roli, 2002). These early methods introduced the idea of abstaining from classification when confidence was low, often using simple model families such as SVM (Cortes & Vapnik, 1995). The phenomenon of neural networks producing overconfident predictions on OOD data was first revealed by Nguyen et al. (2015), highlighting the need for robust detection mechanisms in modern deep learning systems. Building on this foundational work, subsequent research has focused on various techniques for detecting OOD samples, which can be broadly categorized as output-based, feature-based, and gradient-based methods.

Output-Based Methods Many OOD detection approaches directly utilize the model's outputs. Maximum softmax probability, often scaled for calibration, is a classic OOD detection metric (Hendrycks & Gimpel, 2016; Guo et al., 2017). Building on this, temperature scaling combined with input perturbations has shown promise in refining the separation between ID and OOD data (Liang et al., 2018). Additionally, logits themselves have been used for OOD detection, with some methods applying metrics such as KL divergence (Hendrycks et al., 2022). Beyond these, energy-based methods compute OOD scores using energy derived from logits (Liu et al., 2020). Refinements such as truncating activations (Sun et al., 2021b; Sun & Li, 2022) or removing dominant singular values (Song et al., 2022; Djurisic et al., 2022) have been proposed to reduce overconfidence. Generalized entropy scores over logits have also emerged as a robust alternative (Liu et al., 2023). Unlike the aforementioned techniques, GROOD achieves robust OOD detection, even for samples near ID boundaries, by combining gradient norms with class prototype-based representations.

Feature-Based Methods The feature space of neural networks has been a rich avenue for OOD detection. Techniques such as Mahalanobis distance from class centroids (Lee et al., 2018; Ren et al., 2021) and Gram matrices of features (Sastry & Oore, 2020) are prominent examples. Additional methods utilize noise prototypes (Huang et al., 2021a), virtual logits (Wang et al., 2022), and nearest neighbor distances (Sun et al., 2022). Modern Hopfield networks (Zhang et al., 2023b) have also been explored for this purpose. Cosine similarity between test samples and class features (Techapanurak et al., 2020; Chen et al., 2020) has gained traction, with some methods proposing the use of singular vectors for enhanced detection (Zaeemzadeh et al., 2021). Our approach extends these ideas by incorporating an artificial OOD prototype into the feature space, creating a novel gradient-based perspective for OOD detection.

Gradient-Based Methods Gradient-based methods have gained attention for their ability to capture additional information beyond intermediate layers or network outputs. The seminal ODIN approach introduced input perturbations guided by gradients to enhance OOD separation (Hsu et al., 2020). Subsequent works explored the use of gradients with respect to network weights to quantify uncertainties (Lee & AlRegib, 2020; Igoe et al., 2022; Huang et al., 2021b). Another direction utilizes Mahalanobis distances between input gradients, combined with self-supervised classifiers, to detect OOD samples (Sun et al., 2021a). GradNorm calculates an OOD score based on the gradient space of the final layer weights (Huang et al., 2021b). Recent methods, like GAIA (Chen et al., 2023), leverage gradient-based attribution abnormalities with respect to the feature space, combining channel-wise features and zero-deflation patterns. In contrast, GROOD uniquely focuses on gradients with respect to an artificial OOD prototype, capturing subtle differences between ID and OOD data. This approach enables GROOD to improve detection performance by leveraging gradient information in conjunction with prototype-based representations.

3 Preliminaries and Notation

We first introduce the foundational problem of OOD detection and establish the notations.

3.1 Context and Notations

Robust deployment of machine learning models in dynamic real-world environments often requires distinguishing between in-distribution (ID) and out-of-distribution (OOD) data to ensure reliability and safety. To formalize this challenge, we consider a supervised classification problem. Let X denote the input space and $Y = \{1, 2, \dots, C\}$ the label space, where each input-output pair (x, y) is sampled from a joint data distribution P_{XY}. The training set $\mathcal{D}_{\text{in}} = \{(x_i, y_i)\}_{i=1}^{n}$ is assumed to be drawn iid from P_{XY}. Let P_X represent the marginal distribution over X. The marginal distribution of the in-distribution data, denoted as P_{in}, is assumed to be sampled from P_X.

The neural network $f : X \to \mathbb{R}^{|Y|}$ is trained on samples from P_{XY} to produce a logit vector, subsequently used for label prediction. The architecture of f is decomposed as:

$$f = f^{\text{clf}} \circ f^{\text{pen}}, \quad \text{where} \quad f^{\text{pen}} = f^{\text{mid}} \circ f^{\text{early}}. \tag{1}$$

Here, f^{early} extracts low-level features, f^{mid} processes mid-level representations, and f^{pen} produces penultimate features. The final classification module f^{clf} outputs the predictions.

3.2 Problem Setting: Out-of-Distribution Detection

When deploying a machine learning model in practice, the classifier should not only be accurate on ID samples but should also identify any OOD inputs as “unknown”.

Formally, OOD detection can be viewed as a binary classification task. During testing, the task is to determine whether a sample $x \in X$ comes from P_{in} (ID) or not (OOD). The decision can be framed as a level set estimation:

$$G_\tau(x) = \begin{cases} \text{ID}, & \text{if } S(x) \leq \tau, \\ \text{OOD}, & \text{if } S(x) > \tau, \end{cases}$$

where $S : X \to \mathbb{R}$ is a score function quantifying the likelihood of a sample belonging to the ID distribution, and τ is a threshold ensuring that a high fraction (e.g., 95%) of ID data is correctly classified.

4 GROOD Methodology

Overview In this section, we introduce our proposed method GRadient-aware Out-Of-Distribution detection (GROOD), a novel framework for distinguishing between ID and OOD samples. To illustrate the

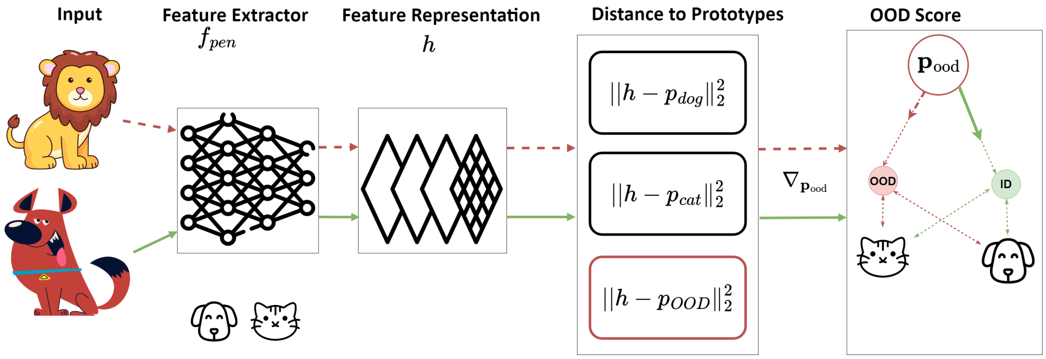


Figure 1: Initially, we build ID class prototypes as the means of the activations of ID data along with an OOD prototype capturing OOD characteristics (§ 5). Subsequently, gradients of the softmax loss built upon the NCPs distance as logits are computed w.r.t. OOD prototype (§ 4.1). Finally, the OOD score is determined using nearest neighbor distance in the gradient space (§ 4.2).

core mechanism, we will initially assume the existence of a representative OOD prototype, the calculation of which will be detailed in § 5.1. Notably, our oracle experiments (§ 5.1) demonstrate the existence of such an OOD prototype that exhibits strong discriminative properties. The framework itself utilizes gradient information with respect to this OOD prototype to capture distributional shifts that may not be apparent in feature representations alone.

The method comprises two primary components, illustrated in Figure 1: (1) A gradient computation framework that quantifies sample responses to an OOD prototype (§ 4.1), and (2) A nearest-neighbor scoring mechanism operating in gradient space (§ 4.2).

4.1 Gradients Computation

Inspired by the neural collapse property (Papyan et al., 2020) and prototype-based recognition (Shu et al., 2019), we construct a distance-based classification framework that incorporates both class and OOD prototypes. The key idea is to transform distances to prototypes into logits, with the crucial addition of an OOD prototype that serves as a reference point for out-of-distribution detection.

For a feature vector h in the penultimate layer space, we define the logit vector as

$$L(h) = -[\|h - \mathbf{p}_1^{\text{pen}}\|_2, \dots, \|h - \mathbf{p}_C^{\text{pen}}\|_2, \|h - \mathbf{p}_{\text{ood}}^{\text{pen}}\|_2], \tag{2}$$

where $\mathbf{p}_i^{\text{pen}}$ represents the prototype for class i, and $\mathbf{p}_{\text{ood}}^{\text{pen}}$ denotes the OOD prototype. These negative distances are transformed into probabilities through the softmax function:

$$p_i(h) = \frac{\exp(L_i(h))}{\sum_{j=1}^{C+1} \exp(L_j(h))}, \quad i = 1, \dots, C+1, \tag{3}$$

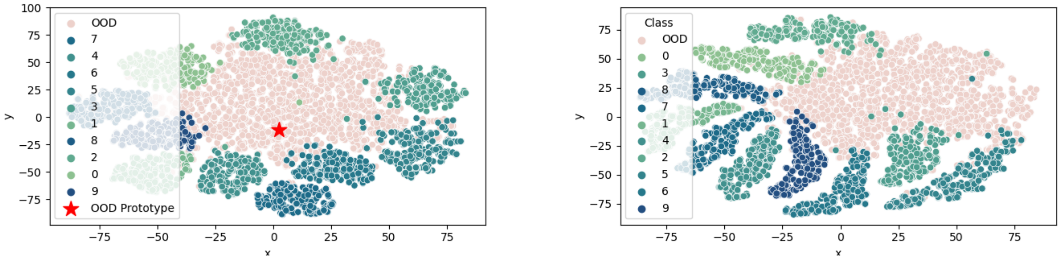
where $p_{C+1}(h) = p_{\text{ood}}(h)$ represents the probability of the sample being OOD.

This formulation enables us to quantify, through the NCP loss, how well a sample aligns with the ID prototypes versus the OOD prototype. For an ID sample, we expect strong alignment with one of the class prototypes and weak alignment with the OOD prototype.

For a given feature vector h and a class $y \in 1, .., C, C+1$, the cross-entropy loss associated with the NCP output $[p_i(h)]_{i=1}^{C+1}$ from equation 3 is given by:

$$H(h, y) = -\log p_y(h), \tag{4}$$

The key insight of our method lies in analyzing the gradient of the loss $H(h, y)$ for some (any) iid class y with respect to the OOD prototype $\mathbf{p}_{\text{ood}}^{\text{pen}}$. Intuitively, this quantity represents the update vector for the



(a) (b)

Figure 2: t-SNE Visualizations of ID and OOD Samples. (a) Distribution within the feature representation space including the OOD prototype using synthetic data. (b) Gradient distributions from a CIFAR-10 pre-trained ResNet-18 model, differentiating ID and OOD samples.

OOD prototype assuming that the feature h corresponds to an iid sample. This gradient can be expressed in closed form as follows:

$$\nabla H(h) := \nabla_{\mathbf{p}_{\text{ood}}^{\text{pen}}} H(h, C+1) = p_{\text{ood}}(h) \frac{h - \mathbf{p}_{\text{ood}}^{\text{pen}}}{\|h - \mathbf{p}_{\text{ood}}^{\text{pen}}\|_2}. \tag{5}$$

These computations are grounded in two core observations. The first is inspired by the Neural Collapse phenomenon (Papyan et al., 2020), which shows that well-trained networks often align penultimate-layer features with their corresponding class prototypes. This alignment justifies using distances in feature space as a reliable signal for class identity. The second is that OOD samples tend to lie outside these tight clusters, often in low-density or dispersed regions. By introducing a synthetic OOD prototype and computing the gradient of prediction confidence with respect to its position, GROOD distinguishes ID from OOD samples by measuring how sensitive each sample's classification confidence is to small shifts in the OOD prototype.

This gradient expression reveals two important properties (see the Appendix for derivation details): 1) The gradient norm equals the softmax probability $p_{\text{ood}}(h)$ of the OOD class, providing a direct measure of out-of-distribution likelihood. 2) The gradient direction, given by the unit vector $(h - \mathbf{p}_{\text{ood}}^{\text{pen}})/\|h - \mathbf{p}_{\text{ood}}^{\text{pen}}\|_2$, captures the geometric relationship between the sample and the OOD prototype.

For ID samples, we observe smaller gradient norms due to lower OOD probabilities. However, the complete gradient vector provides richer information than the norm alone, encoding both magnitude and directional differences between ID and OOD samples. This allows us to detect OOD samples through both the size of the hypothetical update to $\mathbf{p}_{\text{ood}}^{\text{pen}}$ and its direction.

To analyze the potential for enhanced separability, fig. 2 visualizes (a) the penultimate feature space and (b) the gradient space using t-SNE. Our intuition is that the gradient $\nabla H(h)$ will yield a more distinct separation between ID and OOD samples compared to the feature space. As seen in fig. 2 (a), ID samples form class-specific clusters, while OOD samples are scatttered in the space. However, the t-SNE plot of the gradient space in fig. 2 (b) reveals a different spatial arrangement, suggesting that the gradient transformation highlights discriminative characteristics for OOD detection, which GROOD leverages through distance computations in this space (eq. (6)). Thus, the single OOD prototype, $p_{\text{ood}}^{\text{pen}}$, serves as a fixed reference point rather than a geometric representative of all OOD diversity. Its role is not to span the OOD distribution, but to anchor gradient-based sensitivity: how much an input's classification would change if "OODness" were perturbed. The resulting gradient vector encodes both magnitude and direction, providing a rich signal to discriminate between ID and OOD samples. As shown in Figure 2, this transformation yields significantly improved ID-OOD separation in gradient space, even when the raw feature space exhibits overlap.

4.2 Final OOD Score Computation

Having established how to compute discriminative gradients with respect to an OOD prototype, we now address a key challenge: how to effectively use these gradients to distinguish between ID and OOD samples.

Our solution leverages the observation that ID samples produce similar gradient patterns, while OOD samples generate distinctly different ones.

4.2.1 Distance-Based Scoring

For a test sample x_{new}, we first compute its feature representation $h(x_{\text{new}})$ and corresponding gradient $\nabla H(h(x_{\text{new}}))$. Our OOD score is then defined as the distance to the nearest training gradient:

$$S(x_{\text{new}}) = \min_{x \in \mathcal{D}_{\text{in}}} \|\nabla H(h(x_{\text{new}})) - \nabla H(h(x))\|_2 \tag{6}$$

This formulation captures our intuition that OOD samples will produce gradients that deviate significantly from those seen during training. The minimum distance provides a natural measure of "outlierness" - the further a gradient is from its nearest training neighbor, the more likely the sample is to be OOD.

Efficient Implementation A naive implementation of nearest neighbor search in high-dimensional gradient space would be computationally prohibitive. We address this challenge using the FAISS library Douze et al. (2024), which provides efficient approximate nearest neighbor search through inverted lists and quantization. This makes our method practical for large-scale applications while maintaining accuracy.

The preceding section outlined the core methodology of GROOD, detailing how gradients with respect to an OOD prototype are computed and subsequently used within a nearest-neighbor scoring mechanism to distinguish ID and OOD samples. A critical element underpinning the effectiveness of this methodology is the choice and construction of the OOD prototype, $\mathbf{p}_{\text{ood}}^{\text{pen}}$. The nature and location of this prototype in the feature space directly influence the direction and magnitude of the computed gradients, thereby impacting the separability of ID and OOD samples in the gradient space. Therefore, the following section delves into the specific strategies we employ for **Prototype Computation** (§ 5), exploring various approaches to define both the class prototypes and, crucially, the OOD prototype. These approaches are designed to yield a $\mathbf{p}_{\text{ood}}^{\text{pen}}$ that optimizes the discriminative power of the gradient-based OOD score introduced in our methodology.

5 Prototype Computation

5.1 Class and OOD Prototypes

The foundation of our method lies in computing class-discriminative prototypes. For each class, we compute prototypes at both early and penultimate layers as the average of feature vectors:

$$p_y^l = \frac{1}{|X_y|} \sum_{x \in X_y} f^l(x), \quad l \in \{\text{early}, \text{pen}\} \tag{7}$$

where X_y is the set of training instances in class y.

Similarly, the OOD prototype is computed as the average of feature vectors from a dataset X_{ood}:

$$p_{ood}^{\text{pen}} = \frac{1}{|X_{ood}|} \sum_{x \in X_{ood}} f^{\text{pen}}(x) \tag{8}$$

The less variability in the representation space of each ID class as per NC(Papyan et al., 2020), the more effective GROOD will be in distinguishing ID and OOD samples.

Clearly, the choice of dataset X_{ood} will have a significant impact on GROOD's ability to differentiate between ID and OOD samples. In the remainder of this section we will first show that, given access to actual OOD data, GROOD performs significantly better than the state of the art (Sec. 5.1.1). Subsequently, we propose a method for synthesizing the OOD data used to calculate the OOD prototype that approximates the performance of the priviliged case (Sec. 5.2).

Table 1: Oracle experiment results compared to SOTA excluding GROOD from Table 2. AUROC (%) on far-OOD and near-OOD detection using 100 prototype samples. Results are averaged over different checkpoints; standard deviations in parentheses.

ID Dataset	**Architecture**	**Local Oracle**		**Global Oracle**		**SOTA**	
		AUROC (%) ↑					
		Far-OOD	Near-OOD	Far-OOD	Near-OOD	Far-OOD	Near-OOD
CIFAR-10	ResNet-18	96.7 (±0.2)	95.4 (±0.1)	94.8 (±0.1)	90.8 (±0.1)	94.7	90.7
CIFAR-100	ResNet-18	94.8 (±0.2)	85.5 (±0.5)	88.1 (±0.03)	82.1 (±0.1)	82.4	81.3
ImageNet-200	ResNet-18	98.8 (±0.2)	92.1 (±0.2)	94.22 (±0.02)	84.1 (±0.2)	93.16	82.9
ImageNet-1K	ResNet-50	97.9	91.0	96.2	79	95.1	78.1

5.1.1 "Oracle" Experiment

To validate the core idea that an OOD prototype can help distinguish ID and OOD data, we performed an "oracle" experiment, assuming temporary access to some OOD information.

Local Oracle (Idealized): For each ID-OOD test pair, we used a small sample (100) from the *test* OOD data to build a specific OOD prototype. We then tested the remaining OOD data. This setup essentially asks: "If we had perfect knowledge of a small subset of the specific OOD data we'd encounter, how well could our method perform?". The remarkable performance achieved in this setting (over 95% AUROC on far-OOD detection, as shown in table 1) highlights the inherent potential of an OOD prototype tailored to the specific distributional shift.

Global Oracle (Generalizable): For each ID dataset, we used a small "validation" portion of *all other* OOD datasets to create a single, general OOD prototype. We then tested on the held-out portion of each specific OOD dataset. This setup aims to mimic a scenario where we have access to some diverse OOD data (the validation sets) but not the specific OOD data we are currently testing on. The results from the Global Oracle provide insights into how well a more general OOD prototype can generalize across different out-of-distribution scenarios.

Despite the strong performance in the far-OOD detection tasks under both oracle settings, the notably lower performance on near-OOD detection underscores the inherent difficulty of distinguishing between distributions that are semantically or statistically close to the in-distribution data. Nevertheless, the oracle experiments strongly suggest that the concept of an OOD prototype holds significant promise for OOD detection when a representative prototype can be effectively determined.

On the other hand, in real-world scenarios, we cannot construct this oracle prototype using test OOD data. This raises a key question: *How can we approximate this optimal OOD prototype without access to the test distribution?*

5.2 Practical OOD Prototype Construction

The oracle experiments presented in § 5.1 demonstrated the significant potential of employing an OOD prototype to distinguish between ID and OOD data, achieving high performance when even approximate knowledge of the OOD distribution was available. This motivates our goal: to effectively approximate such an optimal OOD prototype, $\mathbf{p}_{\text{ood}}^{\text{pen}}$, without requiring access to the specific test OOD distribution, which is unavailable in practical scenarios.

While real-world OOD samples exhibit considerable diversity, representing them with a single prototype $\mathbf{p}_{\text{ood}}^{\text{pen}}$ proves effective within our gradient-aware framework. Our approach does not aim to represent all OOD samples geometrically, but rather uses the prototype as a fixed reference point. The core OOD score relies on the sensitivity of the NCP loss to this prototype, measured via the gradient $\nabla_{\mathbf{p}_{\text{ood}}^{\text{pen}}} H(h)$ (§ 4, eq. (5)). We hypothesize that OOD samples, inherently deviating from learned ID manifolds, exhibit distinct gradient sensitivity patterns (both magnitude and direction) relative to this OOD reference point, allowing separation from more stable ID samples.

We propose several complementary approaches to construct $\mathbf{p}_{\text{ood}}^{\text{pen}}$, leveraging information from an auxiliary OOD dataset X_{ood}:

Synthetic OOD Generation using mixup Our first approach requires no external OOD data, instead synthesizing OOD-like features by exploiting decision boundaries. We perform guided prototype interpolation towards the second-highest predicted class c_2 at an early layer (after the first block):

$$\hat{h}(x) = f^{\text{mid}}\left(\lambda f^{\text{early}}(x) + (1-\lambda)\mathbf{p}_{c_2}^{\text{early}}\right) \tag{9}$$

where $\lambda = 0.5$ positions the synthetic samples near decision boundaries. This approach leverages our observation that early layer representations are more sensitive to perturbations, making them ideal for generating OOD-like features.

Auxiliary OOD Validation When available, we can utilize a small auxiliary OOD validation set as X_{ood} to construct the prototype following eq. (8) using 100 OOD validation samples. Importantly, we ensure these samples have no category overlap with the test set. Our method shows remarkable stability to the specific choice of validation samples, with a maximum AUROC standard deviation of only 0.5% across five different random selections.

Proximity-Based OOD Filtering (Postprocessing) Initially, we explored constructing the OOD prototype by simply averaging feature vectors from all available OOD samples. However, this approach yielded prototypes that lacked sufficient discriminative power, resulting in poor separation between ID and OOD data. To address this, we introduce a proximity-based filtering step to refine the OOD prototype, enhancing its ability to distinguish OOD samples from ID samples.

Specifically, given a set of candidate OOD feature vectors, we discard OOD samples whose distance $d_i = \min_j \|f^{\text{pen}}(o_i) - p_j\|_2$ falls below an adaptive threshold τ, computed as the q-th quantile of $\{d_i\}_{i=1}^{n_{\text{ood}}}$. This filtering step refines the OOD prototype by ensuring separation from ID data while preserving representativeness.

A comparison of alternative X_{OOD} generation methods is presented in the appendix. The results reported in the remainder of this paper utilize the OOD prototype derived from synthetic data generated from ID samples.

6 Experiments

For a comprehensive evaluation of GROOD's performance, we adhere to the OpenOOD v1.5 criteria Zhang et al. (2023a); Yang et al. (2022). Results are aggregated in table 2 including the performance of the nearest class prototype used instead of the classification head. For robustness, each evaluation metric except for ImageNet-1k is derived from three runs with unique initialization seeds. In the case of ImageNet-1k, we report results based on a single seed run provided by torchvision maintainers & contributors (2016).

Experimental Setup We evaluate performance using the Area Under the Receiver Operating Characteristic curve (AUROC), where higher values are better. Our benchmarking strategy follows the OpenOOD framework Zhang et al. (2023a); Yang et al. (2022), involving four core ID datasets (CIFAR-10, CIFAR-100, ImageNet-200, ImageNet-1k) and examining both near and far-OOD scenarios. For CIFAR-10/100 (50k train/10k test images each), near-OOD datasets are CIFAR-100/TinyImageNet, and far-OOD are MNIST, SVHN, Textures, and Places365. For ImageNet-200 (200 classes, 64x64 resolution), near-OOD datasets are SSB-hard/NINCO, and far-OOD are iNaturalist, Textures, and OpenImage-O; ImageNet-1k shares these OOD datasets. Regarding configuration, we deploy ResNet-18 for CIFAR-10/100 and ImageNet-200 using pre-trained checkpoints from OpenOOD for consistency, testing with three distinct seeds for robustness. For ImageNet-1k, we apply pre-trained torchvision models (ResNet-50, ViT-B-16, Swin-T) to explore GROOD's effectiveness in a broader context than OpenOOD v1 Yang et al. (2022). To allow for reproducibility and facilitate further research, the complete code, including training and evaluation scripts, is available at: `https://anonymous.4open.science/r/grood`.

For CIFAR-10/100 and ImageNet-200, we train ResNet-18 models for 100 epochs using SGD with momentum 0.9, weight decay 5e-4, cosine learning rate decay (starting from 0.1), and batch sizes of 128 (CIFAR) and 256 (ImageNet-200). For ImageNet-1K, we use pretrained models from `torchvision` (ResNet-50, ViT, Swin). When fine-tuning is required, we follow the OpenOOD v1.5 protocol (Zhang et al., 2023a) with 30 epochs, learning rate 0.001, and batch size 256.

Table 2: Main results from OpenOOD v1.5 on standard OOD detection (AUROC). GROOD using synthetic OOD data(§ 5.2) shows superior results compared to existing baselines.

	CIFAR-10		**CIFAR-100**		**ImageNet-200**		**ImageNet-1K**	
ID Acc. (%)	95.06%$_{(\pm0.30)}$		77.25%$_{(\pm0.10)}$		86.37%$_{(\pm0.08)}$		76.18%	
NCP Acc. (%)	95.01%$_{(\pm0.08)}$		77.10%$_{(\pm0.001)}$		85.75%$_{(\pm0.003)}$		71.38%	
Method	Near-OOD (%) ↑	Far-OOD (%) ↑	Near-OOD (%) ↑	Far-OOD (%) ↑	Near-OOD (%) ↑	Far-OOD (%) ↑	Near-OOD (%) ↑	Far-OOD (%) ↑
OpenMax Bendale & Boult (2015)	87.62%$_{(\pm0.29)}$	89.62%$_{(\pm0.19)}$	76.41%$_{(\pm0.25)}$	79.48%$_{(\pm0.41)}$	80.27%$_{(\pm0.10)}$	90.20%$_{(\pm0.17)}$	74.77%	89.26%
MSP Hendrycks & Gimpel (2016)	88.03%$_{(\pm0.25)}$	90.73%$_{(\pm0.43)}$	80.27%$_{(\pm0.11)}$	77.76%$_{(\pm0.44)}$	83.34%$_{(\pm0.06)}$	90.13%$_{(\pm0.09)}$	76.02%	85.23%
ODIN Liang et al. (2018)	82.87%$_{(\pm1.85)}$	87.96%$_{(\pm0.61)}$	79.90%$_{(\pm0.11)}$	79.28%$_{(\pm0.21)}$	80.27%$_{(\pm0.08)}$	91.71%$_{(\pm0.19)}$	74.75%	89.47%
MDS Lee et al. (2018)	84.20%$_{(\pm2.40)}$	89.72%$_{(\pm1.36)}$	58.69%$_{(\pm0.09)}$	69.39%$_{(\pm1.39)}$	61.93%$_{(\pm0.51)}$	74.72%$_{(\pm0.26)}$	55.44%	74.25%
EBO Liu et al. (2020)	87.58%$_{(\pm0.46)}$	91.21%$_{(\pm0.92)}$	80.91%$_{(\pm0.08)}$	79.77%$_{(\pm0.61)}$	82.50%$_{(\pm0.05)}$	90.86%$_{(\pm0.21)}$	75.89%	89.47%
ReAct Sun et al. (2021b)	87.11%$_{(\pm0.61)}$	90.42%$_{(\pm1.41)}$	80.77%$_{(\pm0.05)}$	80.39%$_{(\pm0.49)}$	81.87%$_{(\pm0.98)}$	92.31%$_{(\pm0.56)}$	77.38%	93.67%
MLS Hendrycks et al. (2022)	87.52%$_{(\pm0.47)}$	91.10%$_{(\pm0.89)}$	81.05%$_{(\pm0.07)}$	79.67%$_{(\pm0.57)}$	82.90%$_{(\pm0.04)}$	91.11%$_{(\pm0.19)}$	76.46%	89.57%
GradNorm Huang et al. (2021b)	54.90%$_{(\pm0.98)}$	57.55%$_{(\pm3.22)}$	70.13%$_{(\pm0.47)}$	69.14%$_{(\pm1.05)}$	72.75%$_{(\pm0.48)}$	84.26%$_{(\pm0.87)}$	72.96%	90.25%
GAIA Chen et al. (2023)	85.1%$_{(\pm10.2)}$	92.1%$_{(\pm2.9)}$	70.75%$_{(\pm2.11)}$	86.2%$_{(\pm5.1)}$	75.1%$_{(\pm9.8)}$	88.14%$_{(\pm1.8)}$	66.98%	90.2%
CIDER Ming et al. (2023)	90.7%$_{(\pm0.1)}$	**94.7%**$_{(\pm0.36)}$	73.10%$_{(\pm0.3)}$	80.49%$_{(\pm0.68)}$	80.58%$_{(\pm1.7)}$	90.66%$_{(\pm1.6)}$	68.9%	92.18%
VIM Wang et al. (2022)	88.68%$_{(\pm0.28)}$	93.48%$_{(\pm0.24)}$	74.98%$_{(\pm0.13)}$	81.70%$_{(\pm0.62)}$	78.68%$_{(\pm0.24)}$	91.26%$_{(\pm0.19)}$	72.08%	92.68%
KNN Sun et al. (2022)	90.64%$_{(\pm0.20)}$	92.96%$_{(\pm0.14)}$	80.18%$_{(\pm0.15)}$	82.40%$_{(\pm0.17)}$	81.57%$_{(\pm0.17)}$	93.16%$_{(\pm0.22)}$	71.10%	90.18%
DICE Sun & Li (2022)	78.34%$_{(\pm0.79)}$	84.23%$_{(\pm1.89)}$	79.38%$_{(\pm0.23)}$	80.01%$_{(\pm0.18)}$	81.78%$_{(\pm0.14)}$	90.80%$_{(\pm0.31)}$	73.07%	90.95%
ASH Djurisic et al. (2022)	75.27%$_{(\pm1.04)}$	78.49%$_{(\pm2.58)}$	78.20%$_{(\pm0.15)}$	80.58%$_{(\pm0.66)}$	82.38%$_{(\pm0.19)}$	93.90%$_{(\pm0.27)}$	78.17%	95.1%
SHE Zhang et al. (2023b)	81.54%$_{(\pm0.51)}$	85.32%$_{(\pm1.43)}$	78.95%$_{(\pm0.18)}$	76.92%$_{(\pm1.16)}$	80.18%$_{(\pm0.25)}$	89.81%$_{(\pm0.61)}$	73.78%	90.92%
GEN Liu et al. (2023)	88.2%$_{(\pm0.3)}$	91.35%$_{(\pm0.55)}$	**81.31%**$_{(\pm0.1)}$	79.68%$_{(\pm0.6)}$	82.9%$_{(\pm0.34)}$	91.36%$_{(\pm0.45)}$	76.85%	89.76%
fdbd Liu & Qin (2023)	90.4%$_{(\pm0.12)}$	93.16%$_{(\pm0.25)}$	81.2%$_{(\pm0.05)}$	79.85%$_{(\pm0.15)}$	**84.2%**$_{(\pm0.3)}$	93.4%$_{(\pm0.2)}$	76.6%	92.7%
NCI Liu & Qin (2025)	88.8%$_{(\pm0.1)}$	91.26%$_{(\pm0.2)}$	81%$_{(\pm0.2)}$	81.3%$_{(\pm0.15)}$	83.5%$_{(\pm0.4)}$	**93.7%**$_{(\pm0.15)}$	78.6%	**95.5%**
GROOD(OURS)	**91.16%**$_{(\pm0.001)}$	93.8%$_{(\pm0.02)}$	78.9%$_{(\pm0.05)}$	**84.44%**$_{(\pm0.9)}$	**83.4%**$_{(\pm0.12)}$	92.19%$_{(\pm0.12)}$	**78.91%**	94.8 %

Main Results Discussion GROOD shows strong performance across datasets, but performance varies depending on the trade-off between Near- and Far-OOD detection. On CIFAR-100, GROOD achieves state-of-the-art Far-OOD AUROC (84.44%) among post-hoc methods and competitive Near-OOD performance (78.9%). While some methods like VIM (Wang et al., 2022) (81.70%) report higher Near-OOD scores, they trade off Far-OOD robustness. ~~GROOD achieves top performance in several categories, notably securing the best results for Near-OOD detection on CIFAR-10 (91.16%), ImageNet-200 (83.4%), and ImageNet-1K (78.91%), as well as for Far-OOD detection on CIFAR-100 (84.44%).~~

On ImageNet-1k, GROOD achieves a top Far-OOD score (94.8%) but a lower Near-OOD score (78.91%), whereas CombOOD (Rajasekaran et al., 2024) yields 95.22% Near-OOD and 90.24% Far-OOD.

These results reflect a key characteristic of GROOD: its effectiveness depends on the structure of the ID feature space. As GROOD relies on geometric separation of class prototypes (inspired by Neural Collapse (Papyan et al., 2020)), its performance can degrade when ID representations are less well-clustered. Additionally, the choice of OOD prototype impacts this trade-off. For example, using an "ID-corrupted val" prototype improves Near-OOD AUROC to 80.27% (CIFAR-100) and 83.5% (ImageNet-1k), while maintaining strong Far-OOD scores.

Importantly, many top-performing methods on the OpenOOD leaderboard require access to OOD data (e.g., OE, CIDER). GROOD remains fully post-hoc and training-free, making it more practical for deployment across varied scenarios.

Performance on Transformer Architectures Further demonstrating the robustness and generalization capabilities of our approach, table 3 presents the OOD detection performance on ImageNet-1K using Transformer-based architectures, ViT-B-16 (Dosovitskiy et al., 2020) and Swin-T (Liu et al., 2021b). GROOD maintains strong performance, achieving the top AUROC scores for both Near-OOD and Far-OOD detection on ViT-B-16 (Dosovitskiy et al., 2020), and the best Near-OOD score on Swin-T (Liu et al., 2021b) while being highly competitive for Far-OOD. This contrasts significantly with several other methods, such as GradNorm and ASH, whose performance severely degrades on these Transformer architectures compared

Table 3: Performance comparison (AUROC %) on ImageNet-1K using different architectures.

	ViT-B-16		Swin-T	
Method	Near-OOD (%) ↑	Far-OOD (%) ↑	Near-OOD (%) ↑	Far-OOD (%) ↑
GROOD (OURS)	**76.47%**	**90.84%**	**76.10%**	88.90%
ReACT Sun et al. (2021b)	69.26%	85.69%	75.64%	88.23%
GradNorm Huang et al. (2021b)	39.28%	41.75%	47.58%	35.47%
KNN Sun et al. (2022)	74.11%	90.60%	71.62%	**89.37%**
ASH Djurisic et al. (2022)	53.21%	51.56%	46.47%	44.64%

to their ResNet-based results. This suggests that GROOD's mechanism, relying on gradient sensitivity relative to class and OOD prototypes, generalizes more effectively across fundamentally different architectural paradigms than methods potentially more sensitive to specific CNN feature properties.

6.1 Ablation Studies

To understand the contribution of each component in GROOD, we conduct a series of ablation studies on CIFAR-10 (table 4). Each study isolates a key design choice and quantifies its impact on both near-OOD and far-OOD detection performance.

Table 4: Ablation study using different losses and OOD scores to show the effectiveness of each proposed part. Evaluation done on CIFAR-10.

Model Variant	**AUROC (%)**	
	Far-OOD	**Near-OOD**
(1) Distance to the Noise prototype	84.3$_{(\pm 6.1)}$	79.9$_{(\pm 6.5)}$
(2) Gradient L1-norm only	92.4$_{(\pm 0.48)}$	89.35$_{(\pm 0.41)}$
(3) Grads. wrt class prototypes	92.7$_{(\pm 0.15)}$	89.9$_{(\pm 0.05)}$
(4) OOD prototype with uniform noise only	91.7$_{(\pm 0.6)}$	88.1$_{(\pm 0.55)}$
GROOD	**93.8**$_{(\pm 0.02)}$	**91.16**$_{(\pm 0.001)}$

Distance vs. Gradient-based Scoring Our first experiment examines whether the gradient computation is truly necessary. We compare directly using the distance to the OOD prototype against our full gradient-based approach. The significant performance gap (79.9% vs. 91.16% for near-OOD and 84.3% vs 93.8% for far-OOD) demonstrates that gradients capture richer information about sample distribution than raw distances alone.

Nearest Neighbor vs. Gradient Norm While prior work like GradNorm Huang et al. (2021b) uses L1-norm of gradients as the OOD score, we hypothesized that our full approach, GROOD, would be more informative. The results support this: GROOD achieves 93.8% AUROC on far-OOD detection and 91.16% on near-OOD detection compared to 92.4% and 89.35% respectively with gradient norm alone, suggesting that the combination of our design choices in GROOD provides valuable signal beyond the magnitude of the gradient alone.

OOD vs. Class Prototypes A natural question is whether we need a dedicated OOD prototype at all - could we achieve similar results using gradients with respect to class prototypes? The experiment shows that OOD-specific prototypes provide superior performance (93.8% vs. 92.7% on far-OOD and 91.16% vs 89.9% on near-OOD), validating our design choice to explicitly model out-of-distribution behavior.

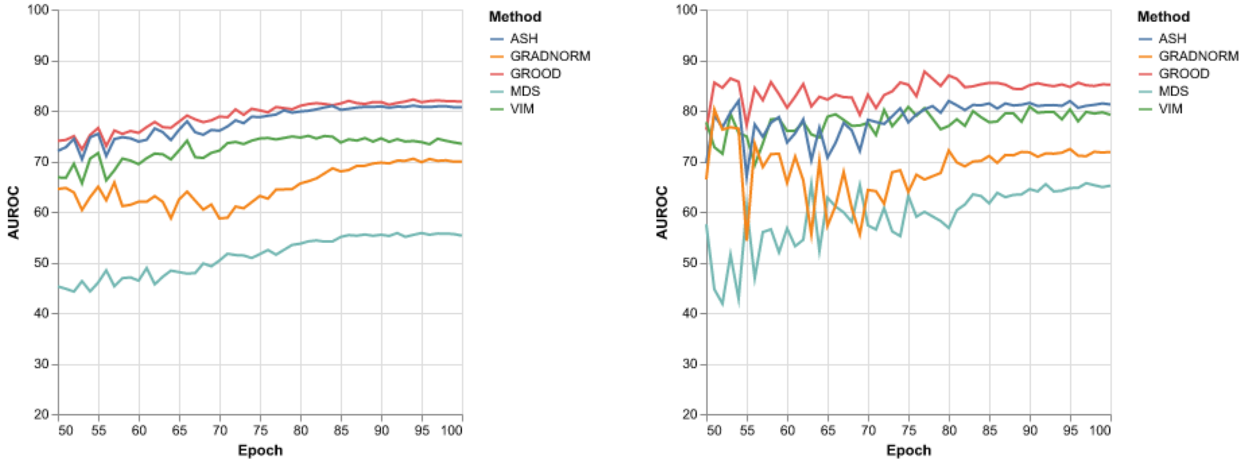
Impact of Noise Sources Finally, we investigate the value of our synthetic OOD data generation for OOD prototype construction compared to simply using uniform noise. Using only uniform noise degrades performance by 2.1% on far-OOD (93.8% vs 91.7%) and 3.06% on near-OOD detection (91.16% vs 88.1%), respectively, demonstrating the benefit of our comprehensive synthetic approach to prototype construction over basic uniform noise.

These ablations collectively validate GROOD's key design choices: each component contributes meaningfully to the final performance, with the full method achieving the best results across all metrics. The consistent improvements and low standard deviations ($\leq 0.6\%$) across experiments indicate the robustness of our approach.

6.2 Robustness to Checkpoint Choice

Table 5: Standard Deviation of AUROC for the Last 15 Epochs on CIFAR-100

Method	Std (Far OOD)	Std (Near OOD)
MDS Lee et al. (2018)	1.05	0.24
ODIN Liang et al. (2018)	0.40	0.22
GradNorm Huang et al. (2021b)	0.62	0.72
VIM Wang et al. (2022)	0.95	0.37
GROOD	**0.38**	**0.2**



(a) Near-OOD AUROC

(b) Far-OOD AUROC

Figure 3: AUROC Performance on Cifar100 (Near and Far OOD) across different checkpoints showing the stability of GROOD

The AUROC metric, while widely used for evaluating OOD detection, can exhibit instability during training, particularly in the later stages. This instability means that small fluctuations in the model's weights can lead to significant variations in AUROC scores, making the selection of an optimal checkpoint challenging. The AUROC curve can vary sharply, even when the test error is relatively stable, indicating a sensitivity to minor weight perturbations. In contrast, GROOD's design contributes to more stable OOD detection performance. The key intuition behind GROOD's robustness lies in its focus on the sensitivity of weights relative to the OOD prototype. Throughout training, while the representation space and the OOD prototype's absolute location change as the network's weights are updated, their inherent relationship—the sensitivity—remains stable, leading to consistent OOD detection , as further evidenced by the reduced standard deviations shown in table 5 and fig. 3.

Further ablation experiments and detailed analyses are provided in the Appendix.

7 Discussion and Conclusion

In this work, we presented GRadient-aware Out-Of-Distribution detection (GROOD), a novel approach to address the critical challenge of detecting OOD samples in DNN-based image classifiers. GROOD leverages the complementary strengths of gradient information and distance metrics, demonstrating a significant advancement in identifying OOD instances, which is crucial for deploying DNNs reliably and safely in real-world applications (Litjens et al., 2017; Bojarski et al., 2016). The extensive experimental evaluation across various benchmarks illustrates the efficacy of GROOD in detecting both near and far OOD samples, exhibiting robust performance independent of specific DNN architectures or training data distributions. Furthermore, GROOD's ability to operate effectively without extensive hyper-parameter tuning underscores its practicality and adaptability in diverse settings.

However, it is important to consider the appropriate use cases and potential limitations of GROOD. As GROOD relies on the concept of class prototypes and is inspired by the Neural Collapse phenomenon (Papyan et al., 2020), it is inherently well-suited for applications where the in-distribution data exhibits a clear cluster structure in the feature space. In such scenarios, GROOD's ability to effectively distinguish between these clusters and identify deviations becomes particularly advantageous. Conversely, GROOD may face challenges in applications where the in-distribution data is inherently noisy, or where the underlying model fails to clearly separates the in-distribution classes.

Beyond application-specific considerations, GROOD also presents certain practical limitations. While our approach demonstrates improved inference speed compared to KNN-based methods, as detailed in appendix A.4, there are computational costs associated with storing and processing gradient information, which can become significant for very large datasets or high-resolution images. Furthermore, the construction of effective OOD prototypes requires careful consideration. As discussed in § 5.2, while GROOD is relatively robust to the specific choice of auxiliary OOD samples, the diversity and representativeness of these samples can influence performance as shown in appendix A.2.

Several subtle but important design choices were made during the development of GROOD. For example, the proximity-based filtering step in OOD prototype construction, as described in § 5.2, proved crucial for enhancing the discriminative power of the prototype. We also observed that using a mixup strategy targeting the **second**-highest predicted class for synthetic OOD data generation led to better results than interpolating towards a random class, likely due to the resulting samples being closer to the decision boundary.

In conclusion, GROOD offers a promising approach for OOD detection, particularly in applications with well-structured in-distribution data. Future research could explore methods to extend GROOD to handle more complex data distributions or to further optimize its computational efficiency.

References

Martín Arjovsky, Léon Bottou, Ishaan Gulrajani, and David Lopez-Paz. Invariant Risk Minimization. *CoRR*, abs/1907.02893, 2019. URL http://arxiv.org/abs/1907.02893.

Abhijit Bendale and Terrance E. Boult. Towards Open Set Deep Networks. *CoRR*, abs/1511.06233, 2015. URL http://arxiv.org/abs/1511.06233.

Mariusz Bojarski, Davide Del Testa, Daniel Dworakowski, Bernhard Firner, Beat Flepp, Prasoon Goyal, Lawrence D. Jackel, Mathew Monfort, Urs Muller, Jiakai Zhang, Xin Zhang, Jake Zhao, and Karol Zieba. End to End Learning for Self-driving Cars. *CoRR*, abs/1604.07316, 2016. URL http://arxiv.org/abs/1604.07316.

Jinggang Chen, Junjie Li, Xiaoyang Qu, Jianzong Wang, Jiguang Wan, and Jing Xiao. GAIA: Delving into Gradient-based Attribution Abnormality for Out-of-distribution Detection. In Alice Oh, Tristan Naumann, Amir Globerson, Kate Saenko, Moritz Hardt, and Sergey Levine (eds.), *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023*, 2023. URL http://papers.nips.cc/paper_files/paper/2023/hash/fcdccd419c4dc471fa3b73ec97b53789-Abstract-Conference.html.

Xingyu Chen, Xuguang Lan, Fuchun Sun, and Nanning Zheng. A Boundary Based Out-of-distribution Classifier for Generalized Zero-shot Learning. *CoRR*, abs/2008.04872, 2020. URL https://arxiv.org/abs/2008.04872.

CK Chow. On optimum recognition error and reject tradeoff. *IEEE Transactions on Information Theory*, 16(1):41–46, 1970.

Corinna Cortes and Vladimir Vapnik. Support-vector Networks. *Mach. Learn.*, 20(3):273–297, 1995. doi: 10.1007/BF00994018. URL https://doi.org/10.1007/BF00994018.

Andrija Djurisic, Nebojsa Bozanic, Arjun Ashok, and Rosanne Liu. Extremely Simple Activation Shaping for Out-of-distribution Detection. *CoRR*, abs/2209.09858, 2022. doi: 10.48550/ARXIV.2209.09858. URL https://doi.org/10.48550/arXiv.2209.09858.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. *CoRR*, abs/2010.11929, 2020. URL https://arxiv.org/abs/2010.11929.

Matthijs Douze, Alexandr Guzhva, Chengqi Deng, Jeff Johnson, Gergely Szilvasy, Pierre-Emmanuel Mazaré, Maria Lomeli, Lucas Hosseini, and Hervé Jégou. The Faiss library. *CoRR*, abs/2401.08281, 2024. doi: 10.48550/ARXIV.2401.08281. URL https://doi.org/10.48550/arXiv.2401.08281.

John C. Duchi and Hongseok Namkoong. Learning Models with Uniform Performance via Distributionally Robust Optimization. *CoRR*, abs/1810.08750, 2018. URL http://arxiv.org/abs/1810.08750.

Giorgio Fumera and Fabio Roli. Support Vector Machines with Embedded Reject Option. In Seong-Whan Lee and Alessandro Verri (eds.), *Pattern Recognition with Support Vector Machines, First International Workshop, SVM 2002, Niagara Falls, Canada, August 10, 2002, Proceedings*, volume 2388 of *Lecture Notes in Computer Science*, pp. 68–82. Springer, 2002. doi: 10.1007/3-540-45665-1_6. URL https://doi.org/10.1007/3-540-45665-1_6.

Yarin Gal and Zoubin Ghahramani. Dropout as a Bayesian Approximation: Representing Model Uncertainty in Deep Learning. *CoRR*, abs/1506.02142, 2015. URL http://arxiv.org/abs/1506.02142.

Ian J. Goodfellow, Yoshua Bengio, and Aaron C. Courville. *Deep Learning.* Adaptive computation and machine learning. MIT Press, 2016. ISBN 978-0-262-03561-3. URL http://www.deeplearningbook.org/.

Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q. Weinberger. On Calibration of Modern Neural Networks. *CoRR*, abs/1706.04599, 2017. URL http://arxiv.org/abs/1706.04599.

Hany Hassan, Mostafa Elaraby, and Ahmed Y. Tawfik. Synthetic Data for Neural Machine Translation of Spoken-dialects. In Sakriani Sakti and Masao Utiyama (eds.), *Proceedings of the 14th International Conference on Spoken Language Translation, IWSLT 2017, Tokyo, Japan, December 14-15, 2017*, pp. 82–89. International Workshop on Spoken Language Translation, 2017. URL https://aclanthology.org/2017.iwslt-1.12.

Dan Hendrycks and Thomas G. Dietterich. Benchmarking Neural Network Robustness to Common Corruptions and Perturbations. In *7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6-9, 2019.* OpenReview.net, 2019. URL https://openreview.net/forum?id=HJz6tiCqYm.

Dan Hendrycks and Kevin Gimpel. A Baseline for Detecting Misclassified and Out-of-distribution Examples in Neural Networks. *CoRR*, abs/1610.02136, 2016. URL http://arxiv.org/abs/1610.02136.

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 Kamalika Chaudhuri, Stefanie Jegelka, Le Song, Csaba Szepesvári, Gang Niu, and Sivan Sabato (eds.), *International Conference on Machine Learning, ICML 2022, 17-23 July 2022, Baltimore, Maryland, USA*, volume 162 of *Proceedings of Machine Learning Research*, pp. 8759–8773. PMLR, 2022. URL https://proceedings.mlr.press/v162/hendrycks22a.html.

Yen-Chang Hsu, Yilin Shen, Hongxia Jin, and Zsolt Kira. Generalized ODIN: Detecting Out-of-distribution Image without Learning from Out-of-distribution Data. *CoRR*, abs/2002.11297, 2020. URL https://arxiv.org/abs/2002.11297.

Haiwen Huang, Zhihan Li, Lulu Wang, Sishuo Chen, Xinyu Zhou, and Bin Dong. Feature Space Singularity for Out-of-distribution Detection. In Huáscar Espinoza, John A. McDermid, Xiaowei Huang, Mauricio Castillo-Effen, Xin Cynthia Chen, José Hernández-Orallo, Seán Ó hÉigeartaigh, and Richard Mallah (eds.), *Proceedings of the Workshop on Artificial Intelligence Safety 2021 (SafeAI 2021) co-located with the Thirty-Fifth AAAI Conference on Artificial Intelligence (AAAI 2021), Virtual, February 8, 2021*, volume 2808 of *CEUR Workshop Proceedings*. CEUR-WS.org, 2021a. URL https://ceur-ws.org/Vol-2808/Paper_7.pdf.

Rui Huang, Andrew Geng, and Yixuan Li. On the Importance of Gradients for Detecting Distributional Shifts in the Wild. *CoRR*, abs/2110.00218, 2021b. URL https://arxiv.org/abs/2110.00218.

Conor Igoe, Youngseog Chung, Ian Char, and Jeff Schneider. How Useful are Gradients for OOD Detection Really? *CoRR*, abs/2205.10439, 2022. doi: 10.48550/ARXIV.2205.10439. URL https://doi.org/10.48550/arXiv.2205.10439.

Konstantin Kirchheim, Marco Filax, and Frank Ortmeier. PyTorch-OOD: A Library for Out-of-distribution Detection based on PyTorch. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops, CVPR Workshops 2022, New Orleans, LA, USA, June 19-20, 2022*, pp. 4350–4359. IEEE, 2022. doi: 10.1109/CVPRW56347.2022.00481. URL https://doi.org/10.1109/CVPRW56347.2022.00481.

Shu Kong and Deva Ramanan. OpenGAN: Open-set Recognition via Open Data Generation. *CoRR*, abs/2104.02939, 2021. URL https://arxiv.org/abs/2104.02939.

Vignesh Kothapalli, Ebrahim Rasromani, and Vasudev Awatramani. Neural Collapse: A Review on Modelling Principles and Generalization. *CoRR*, abs/2206.04041, 2022. doi: 10.48550/ARXIV.2206.04041. URL https://doi.org/10.48550/arXiv.2206.04041.

Alex Krizhevsky, Ilya Sutskever, and Geoffrey E. Hinton. ImageNet Classification with Deep Convolutional Neural Networks. In Peter L. Bartlett, Fernando C. N. Pereira, Christopher J. C. Burges, Léon Bottou, and Kilian Q. Weinberger (eds.), *Advances in Neural Information Processing Systems 25: 26th Annual Conference on Neural Information Processing Systems 2012. Proceedings of a meeting held December 3-6, 2012, Lake Tahoe, Nevada, United States*, pp. 1106–1114, 2012. URL https://proceedings.neurips.cc/paper/2012/hash/c399862d3b9d6b76c8436e924a68c45b-Abstract.html.

Yann LeCun, Yoshua Bengio, and Geoffrey E. Hinton. Deep learning. *Nat.*, 521(7553):436–444, 2015. doi: 10.1038/NATURE14539. URL https://doi.org/10.1038/nature14539.

Jinsol Lee and Ghassan AlRegib. Gradients as a Measure of Uncertainty in Neural Networks. *CoRR*, abs/2008.08030, 2020. URL https://arxiv.org/abs/2008.08030.

Jinsol Lee, Mohit Prabhushankar, and Ghassan AlRegib. Gradient-based Adversarial and Out-of-distribution Detection. *CoRR*, abs/2206.08255, 2022. doi: 10.48550/ARXIV.2206.08255. URL https://doi.org/10.48550/arXiv.2206.08255.

Kimin Lee, Kibok Lee, Honglak Lee, and Jinwoo Shin. A Simple Unified Framework for Detecting Out-of-distribution Samples and Adversarial Attacks. *CoRR*, abs/1807.03888:184–191, 2018. URL http://arxiv.org/abs/1807.03888.

Shiyu Liang, Yixuan Li, and R. Srikant. Enhancing The Reliability of Out-of-distribution Image Detection in Neural Networks. In *6th International Conference on Learning Representations, ICLR 2018, Vancouver, BC, Canada, April 30 - May 3, 2018, Conference Track Proceedings*. OpenReview.net, 2018. URL https://openreview.net/forum?id=H1VGkIxRZ.

Geert Litjens, Thijs Kooi, Babak Ehteshami Bejnordi, Arnaud Arindra Adiyoso Setio, Francesco Ciompi, Mohsen Ghafoorian, Jeroen A. W. M. van der Laak, Bram van Ginneken, and Clara I. Sánchez. A Survey on Deep Learning in Medical Image Analysis. *CoRR*, abs/1702.05747, 2017. URL `http://arxiv.org/abs/1702.05747`.

Jiashuo Liu, Zheyuan Hu, Peng Cui, Bo Li, and Zheyan Shen. Heterogeneous Risk Minimization. *CoRR*, abs/2105.03818, 2021a. URL `https://arxiv.org/abs/2105.03818`.

Litian Liu and Yao Qin. Fast decision boundary based out-of-distribution detector. *arXiv preprint arXiv:2312.11536*, 2023.

Litian Liu and Yao Qin. Detecting out-of-distribution through the lens of neural collapse. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 15424–15433, 2025.

Weitang Liu, Xiaoyun Wang, John D. Owens, and Yixuan Li. Energy-based Out-of-distribution Detection. *CoRR*, abs/2010.03759, 2020. URL `https://arxiv.org/abs/2010.03759`.

Xixi Liu, Yaroslava Lochman, and Christopher Zach. GEN: Pushing the Limits of Softmax-based Out-of-distribution Detection. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2023, Vancouver, BC, Canada, June 17-24, 2023*, pp. 23946–23955. IEEE, 2023. doi: 10.1109/CVPR52729.2023.02293. URL `https://doi.org/10.1109/CVPR52729.2023.02293`.

Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin Transformer: Hierarchical Vision Transformer using Shifted Windows. *CoRR*, abs/2103.14030, 2021b. URL `https://arxiv.org/abs/2103.14030`.

TorchVision maintainers and contributors. Torchvision: Pytorch's computer vision library. `https://github.com/pytorch/vision`, 2016.

Tomás Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. Efficient Estimation of Word Representations in Vector Space. In Yoshua Bengio and Yann LeCun (eds.), *1st International Conference on Learning Representations, ICLR 2013, Scottsdale, Arizona, USA, May 2-4, 2013, Workshop Track Proceedings*, 2013. URL `http://arxiv.org/abs/1301.3781`.

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, ICLR 2023, Kigali, Rwanda, May 1-5, 2023*. OpenReview.net, 2023. URL `https://openreview.net/forum?id=aEFaE0W5pAd`.

Anh Mai Nguyen, Jason Yosinski, and Jeff Clune. Deep neural networks are easily fooled: High confidence predictions for unrecognizable images. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2015, Boston, MA, USA, June 7-12, 2015*, pp. 427–436. IEEE Computer Society, 2015. doi: 10.1109/CVPR.2015.7298640. URL `https://doi.org/10.1109/CVPR.2015.7298640`.

Paul Novello, Yannick Prudent, Corentin Friedrich, Yann Pequignot, and Matthieu Le Goff. OODEEL, a simple, compact, and hackable post-hoc deep OOD detection for already trained tensorflow or pytorch image classifiers. `https://github.com/deel-ai/oodeel`, 2023.

Vardan Papyan, X. Y. Han, and David L. Donoho. Prevalence of Neural Collapse during the terminal phase of deep learning training. *CoRR*, abs/2008.08186, 2020. URL `https://arxiv.org/abs/2008.08186`.

Magesh Rajasekaran, Md Saiful Islam Sajol, Frej Berglind, Supratik Mukhopadhyay, and Kamalika Das. Combood: A semiparametric approach for detecting out-of-distribution data for image classification. In *Proceedings of the 2024 SIAM International Conference on Data Mining (SDM)*, pp. 643–651. SIAM, 2024.

Jie Ren, Stanislav Fort, Jeremiah Z. Liu, Abhijit Guha Roy, Shreyas Padhy, and Balaji Lakshminarayanan. A Simple Fix to Mahalanobis Distance for Improving Near-OOD Detection. *CoRR*, abs/2106.09022, 2021. URL `https://arxiv.org/abs/2106.09022`.

Chandramouli Shama Sastry and Sageev Oore. Detecting Out-of-distribution Examples with Gram Matrices. In *Proceedings of the 37th International Conference on Machine Learning, ICML 2020, 13-18 July 2020, Virtual Event*, volume 119 of *Proceedings of Machine Learning Research*, pp. 8491–8501. PMLR, 2020. URL http://proceedings.mlr.press/v119/sastry20a.html.

Zheyan Shen, Peng Cui, Tong Zhang, and Kun Kuang. Stable Learning via Sample Reweighting. *CoRR*, abs/1911.12580, 2019. URL http://arxiv.org/abs/1911.12580.

Yu Shu, Yemin Shi, Yaowei Wang, Tiejun Huang, and Yonghong Tian. P-ODN: Prototype based Open Deep Network for Open Set Recognition. *CoRR*, abs/1905.01851, 2019. URL http://arxiv.org/abs/1905.01851.

Karen Simonyan and Andrew Zisserman. Very Deep Convolutional Networks for Large-scale Image Recognition. In Yoshua Bengio and Yann LeCun (eds.), *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*, 2015. URL http://arxiv.org/abs/1409.1556.

Yue Song, Nicu Sebe, and Wei Wang. RankFeat: Rank-1 Feature Removal for Out-of-distribution Detection. *CoRR*, abs/2209.08590, 2022. doi: 10.48550/ARXIV.2209.08590. URL https://doi.org/10.48550/arXiv.2209.08590.

Jingbo Sun, Li Yang, Jiaxin Zhang, Frank Liu, Mahantesh Halappanavar, Deliang Fan, and Yu Cao. Gradient-based Novelty Detection Boosted by Self-supervised Binary Classification. *CoRR*, abs/2112.09815, 2021a. URL https://arxiv.org/abs/2112.09815.

Yiyou Sun and Yixuan Li. DICE: Leveraging Sparsification for Out-of-distribution Detection. In Shai Avidan, Gabriel J. Brostow, Moustapha Cissé, Giovanni Maria Farinella, and Tal Hassner (eds.), *Computer Vision - ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23-27, 2022, Proceedings, Part XXIV*, volume 13684 of *Lecture Notes in Computer Science*, pp. 691–708. Springer, 2022. doi: 10.1007/978-3-031-20053-3_40. URL https://doi.org/10.1007/978-3-031-20053-3_40.

Yiyou Sun, Chuan Guo, and Yixuan Li. ReAct: Out-of-distribution Detection With Rectified Activations. *CoRR*, abs/2111.12797, 2021b. URL https://arxiv.org/abs/2111.12797.

Yiyou Sun, Yifei Ming, Xiaojin Zhu, and Yixuan Li. Out-of-distribution Detection with Deep Nearest Neighbors. *CoRR*, abs/2204.06507, 2022. doi: 10.48550/ARXIV.2204.06507. URL https://doi.org/10.48550/arXiv.2204.06507.

Engkarat Techapanurak, Masanori Suganuma, and Takayuki Okatani. Hyperparameter-free Out-of-distribution Detection Using Cosine Similarity. In Hiroshi Ishikawa, Cheng-Lin Liu, Tomás Pajdla, and Jianbo Shi (eds.), *Computer Vision - ACCV 2020 - 15th Asian Conference on Computer Vision, Kyoto, Japan, November 30 - December 4, 2020, Revised Selected Papers, Part IV*, volume 12625 of *Lecture Notes in Computer Science*, pp. 53–69. Springer, 2020. doi: 10.1007/978-3-030-69538-5_4. URL https://doi.org/10.1007/978-3-030-69538-5_4.

Vikas Verma, Alex Lamb, Christopher Beckham, Amir Najafi, Ioannis Mitliagkas, David Lopez-Paz, and Yoshua Bengio. Manifold Mixup: Better Representations by Interpolating Hidden States. In Kamalika Chaudhuri and Ruslan Salakhutdinov (eds.), *Proceedings of the 36th International Conference on Machine Learning, ICML 2019, 9-15 June 2019, Long Beach, California, USA*, volume 97 of *Proceedings of Machine Learning Research*, pp. 6438–6447. PMLR, 2019. URL http://proceedings.mlr.press/v97/verma19a.html.

Haoqi Wang, Zhizhong Li, Litong Feng, and Wayne Zhang. ViM: Out-Of-distribution with Virtual-logit Matching. *CoRR*, abs/2203.10807, 2022. doi: 10.48550/ARXIV.2203.10807. URL https://doi.org/10.48550/arXiv.2203.10807.

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. In *Thirty-sixth Conference on Neural Information Processing Systems Datasets and Benchmarks Track*, 2022. URL `https://openreview.net/forum?id=gT6j4_tskUt`.

Alireza Zaeemzadeh, Niccoló Bisagno, Zeno Sambugaro, Nicola Conci, Nazanin Rahnavard, and Mubarak Shah. Out-of-distribution Detection Using Union of 1-Dimensional Subspaces. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2021, virtual, June 19-25, 2021*, pp. 9452–9461. Computer Vision Foundation / IEEE, 2021. doi: 10.1109/CVPR46437.2021.00933. URL `https://openaccess.thecvf.com/content/CVPR2021/html/Zaeemzadeh_Out-of-Distribution_Detection_Using_Union_of_1-Dimensional_Subspaces_CVPR_2021_paper.html`.

Jingyang Zhang, Jingkang Yang, Pengyun Wang, Haoqi Wang, Yueqian Lin, Haoran Zhang, Yiyou Sun, Xuefeng Du, Kaiyang Zhou, Wayne Zhang, Yixuan Li, Ziwei Liu, Yiran Chen, and Hai Li. OpenOOD v1.5: Enhanced Benchmark for Out-of-distribution Detection. *CoRR*, abs/2306.09301, 2023a. doi: 10.48550/ARXIV.2306.09301. URL `https://doi.org/10.48550/arXiv.2306.09301`.

Jinsong Zhang, Qiang Fu, Xu Chen, Lun Du, Zelin Li, Gang Wang, Xiaoguang Liu, Shi Han, and Dongmei Zhang. Out-of-distribution Detection based on In-distribution Data Patterns Memorization with Modern Hopfield Energy. In *The Eleventh International Conference on Learning Representations, ICLR 2023, Kigali, Rwanda, May 1-5, 2023*. OpenReview.net, 2023b. URL `https://openreview.net/pdf?id=KkazG4lgKL`.

A Appendix

A.1 GROOD Algorithm

For a comprehensive overview, the complete GROOD algorithm is outlined in algorithm 1 and algorithm 2.

Algorithm 1 GROOD initialization: Compute prototypes and gradients for the training set

```
Require: Training set $\mathcal{D}_{\text{in}}$, trained model $f$
Require: mixup parameter $\lambda$
Compute class prototypes $\mathbf{P}_{\text{early}}$ and $\mathbf{P}_{\text{pen}}$                ▷ eq. (7)
Compute OOD prototype $\mathbf{p}^{\text{pen}}_{\text{ood}}$ using synthetic data generation§ 5.2        ▷ eq. (8)
function COMP_GRAD($h, \mathbf{p}^{\text{pen}}_{y=1,\cdots,C}, \mathbf{p}^{\text{pen}}_{\text{ood}}$)
    compute $\nabla H(h)$                ▷ eq. (5) using $\mathbf{p}^{\text{pen}}_{y=1,\cdots,C}, \mathbf{p}^{\text{pen}}_{\text{ood}}$
    return $\nabla H(h)$
end function
for each $x \in \mathcal{D}_{\text{in}}$ do
    $\nabla(x) =$ COMP_GRAD $(h(x), \mathbf{p}^{\text{pen}}_{y=1,\cdots,C}, \mathbf{p}^{\text{pen}}_{\text{ood}})$
end for
return $\{\nabla(x)\}_{x\in\mathcal{D}_{\text{in}}}, \mathbf{p}^{\text{early}}_{y=1,\cdots,C}, \mathbf{p}^{\text{pen}}_{y=1,\cdots,C}, \mathbf{p}^{\text{pen}}_{\text{ood}}$
```

Algorithm 2 OOD score using GROOD

```
Require: Training dataset $\mathcal{D}_{\text{in}}$, trained model $f$
Require: $\{\nabla(x))\}_{x\in\mathcal{D}_{\text{in}}}, \mathbf{p}^{\text{pen}}_{y=1,\cdots,C}, \mathbf{p}^{\text{pen}}_{\text{ood}}$ from GROOD initialization
Require: function COMP_GRAD (in Alg. 1)
Require: Sample $x_{\text{new}}$, threshold $\tau$
$\nabla(x_{\text{new}}) =$ COMP_GRAD$(h(x_{\text{new}}), \mathbf{p}^{\text{pen}}_{y=1,\cdots,C}, \mathbf{p}^{\text{pen}}_{\text{ood}})$
Compute OOD score using Nearest Neighbor search: $S(x_{\text{new}}) = \min_{x\in\mathcal{D}_{\text{in}}} \|\nabla(x_{\text{new}}) - \nabla(x)\|_2$
return ID if $S(x_{\text{new}}) \leq \tau$ else OOD
```

Table 6: OOD Detection Performance (AUROC %) by Dataset and OOD Prototype Construction Method.

	Cifar-10		**Cifar-100**		**ImageNet-200**		**ImageNet**	
OOD Prototype	Near-OOD	Far-OOD	Near-OOD	Far-OOD	Near-OOD	Far-OOD	Near-OOD	Far-OOD
Synthetic OOD	91.16	93.8	78.9	84.44	83.4	92.19	78.91	94.8
ID-Corrupted Val	90.55	93.88	80.27	81.41	83.9	92.58	83.5	94.6
OpenOOD Val	91.01	94.18	80.92	80.7	81.86	94.77	78.05	96.16
Uniform	90.9	94.1	77.26	84.5	82.84	94.46	75.23	94.55
Mean of Prototypes	88.5	91.4	77.2	81.4	82.3	92.9	71.25	82.21

A.2 Choice of OOD data for OOD prototype computation

OpenOOD Val To form $\mathbf{p}_{\text{ood}}^{\text{pen}}$, we selected 100 data points from an auxiliary OOD validation dataset, as per the OpenOOD framework Zhang et al. (2023a); Yang et al. (2022), ensuring no category overlap with test set images. This selection criterion aligns with practices in established post-hoc analyses Lee et al. (2022; 2018); Kong & Ramanan (2021). Our method demonstrated robustness to the specific choice of OOD samples. An investigation with five distinct sets of 100 OOD validation samples each revealed negligible variation in AUROC, with a maximum standard deviation of **0.5%**, underscoring our approach's stability across different OOD selections.

ID-Corrupted Val We further validated our approach using 100 i.i.d. samples from CIFAR-10-C Hendrycks & Dietterich (2019) for CIFAR-10 as ID and CIFAR-100C for rest of ID datasets including CIFAR-100, ImageNet-200 and ImageNet to ensure no possible overlap or leaks to the test set.

Uniform We try to approximate the representation of OOD data by leveraging uniform noise data. Initially, a batch of random noise images is created using uniformly distributed pixel values across all channels. These noise images are then passed through a neural network to extract logits and features from intermediate layers. An energy score is computed for each image, where lower scores indicate a higher likelihood of being OOD. The images with the lowest energy scores, which are most similar to out-of-distribution (OOD) data, are selected, and their penultimate layer features are extracted. These features are then aggregated to form an OOD prototype.

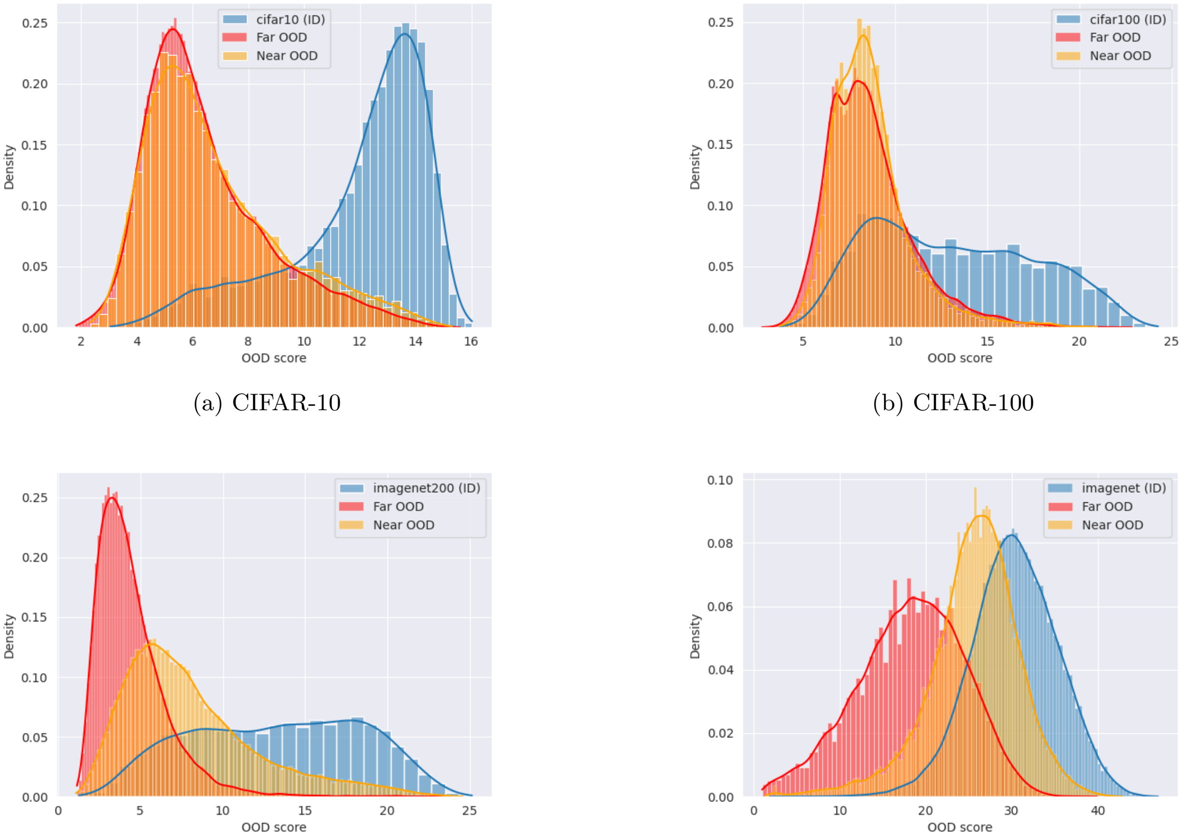
Synthetic OOD To simulate OOD data representations, we utilize a manifold mixup technique on the early layer, similar to the targeted mixup approach described in § 5. However, our method differs in the interpolation target. Instead of interpolating towards the predicted class prototype, we interpolate towards the second-highest predicted class c_2, which is the closest incorrect class on the decision boundary.

Mean of Prototypes Instead of trying to approximate the representation of OOD data using auxiliary OOD data we rely on the mean of ID prototypes.

Table 6 illustrates our method's robustness to different validation OOD datasets.

A.3 Density Plots

To comprehensively evaluate the GROOD method's ability to distinguish between in-distribution (ID) and out-of-distribution (OOD) data, we visualize the distribution of OOD scores across a range of datasets with varying characteristics. Figure 4 presents these visualizations for ID, Near-OOD, and Far-OOD samples on CIFAR-10, CIFAR-100 (datasets with relatively small, natural images), ImageNet-200 (a subset of ImageNet), and ImageNet-1k (a large-scale, complex dataset). Analyzing performance across this spectrum demonstrates GROOD's robustness to differences in image complexity and dataset size. In each subplot, we use density plots to represent the distribution of OOD scores, allowing for a clear visual comparison of separation between ID and OOD distributions.



(a) CIFAR-10

(b) CIFAR-100

(c) ImageNet-200

(d) ImageNet-1k

Figure 4: Distribution of OOD scores on Near-OOD and Far-OOD across different datasets.

A.4 Inference Speed

GROOD introduces a novel Out-of-Distribution (OOD) detection method involving two forward passes for the mixup part which is inexpensive to compute, a backward pass over the OOD prototype which can be computed using its closed form expression as in eq. (5) of the main paper and the ne arest neighbor search which is more computationally intensive. For the latter, Our approach employs the FAISS IndexIVF method for efficient distance computation, utilizing centroids and inverted lists instead of the complete dataset. This technique notably enhances inference speed compared to KNN, particularly in our CIFAR benchmarks. Specifically, on CIFAR-10 and CIFAR-100 datasets, GROOD recorded evaluation inference times over all OOD test sets of **130** seconds and **155** seconds, respectively. This is significantly faster than KNN, which took 434 seconds for CIFAR-10 and 641 seconds for CIFAR-100, demonstrating the efficiency of our approach.

A.5 Ablate value of kth nearest neighbor

Figure 5 illustrates the impact of using the distance to the k-th nearest neighbor, as proposed by Sun et al. (2022). The plot demonstrates that employing the distance to the nearest point in gradient space leads to optimal results.

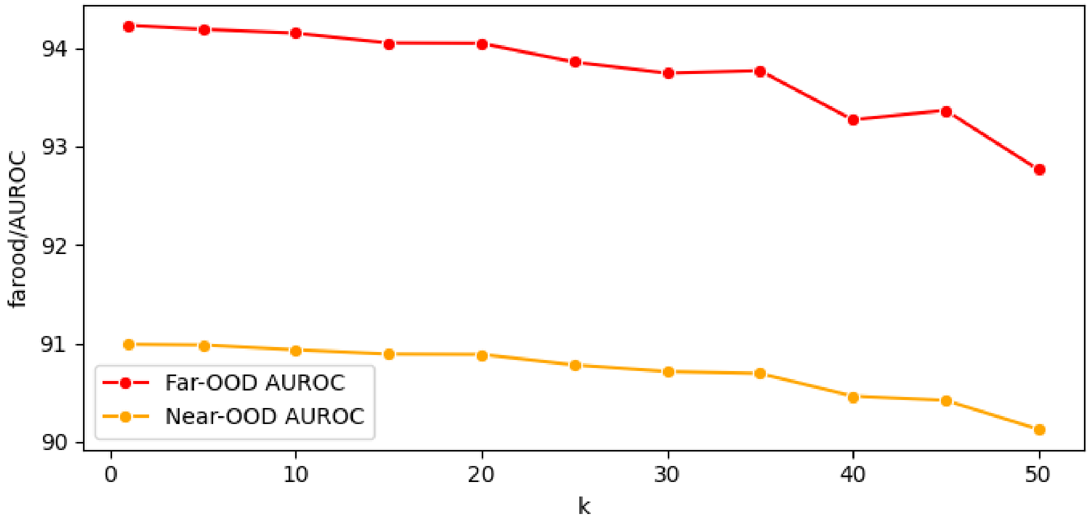


Figure 5: Ablating different values for k-th nearest neighbor parameter on CIFAR-10

A.6 Gradient computation in closed form

This section details the derivation of the closed-form expression for the gradient, as presented in eq. (5) of the main paper. The cross-entropy loss for a logit $L = (L_j)_{j=1}^{C} + 1$ and some (any) ID label y is given by:

$$H(L, y) = -\ln \frac{\exp(L_y)}{\sum_{i=1}^{C+1} \exp(L_j)}$$

The partial derivative of this loss with respect to L_{C+1} is given by:

$$\frac{\partial}{\partial L_{C+1}} H(L, C+1) = \frac{\exp(L_{C+1})}{\sum_{i=1}^{C+1} \exp(L_j)}$$

which is equal to the softmax probability corresponding to the OOD class and does not depend on the specific ID class label y anymore. Now for a feature vector h, the corresponding logit vector $L(h)$ is given by eq. (2). Since $L_{C+1}(h) = \|h - \mathbf{p}_{\text{ood}}^{\text{pen}}\|_2$ is the only logit depending on $\mathbf{p}_{\text{ood}}^{\text{pen}}$, the gradient of the above loss with respect to $\mathbf{p}_{\text{ood}}^{\text{pen}}$ is given by the chain rule:

$$\nabla_{\mathbf{p}_{\text{ood}}^{\text{pen}}} H(L(h), y) = \frac{\partial}{\partial L} H(L(h), y) \nabla_{\mathbf{p}_{\text{ood}}^{\text{pen}}} L(h) = \frac{\exp(L_{C+1})}{\sum_{i=1}^{C+1} \exp(L_j)} \frac{h - \mathbf{p}_{\text{ood}}^{\text{pen}}}{\|h - \mathbf{p}_{\text{ood}}^{\text{pen}}\|_2} = p_{\text{ood}}(h) \frac{h - \mathbf{p}_{\text{ood}}^{\text{pen}}}{\|h - \mathbf{p}_{\text{ood}}^{\text{pen}}\|_2},$$

as desired.

A.7 Impact of Mixup-Trained Backbones on GROOD

Method	Near-OOD AUROC (%)	Far-OOD AUROC (%)
GROOD (mean prototype)	**81.05**	**80.26**
ASH (Djurisic et al., 2022)	79.1	56.0
KNN (Sun et al., 2022)	78.0	**81.85**
GradNorm (Huang et al., 2021b)	50.0	50.0

Table 7: GROOD with mean prototype on manifold mixup-trained ResNet-18.

We investigate how GROOD's OOD prototype construction interacts with backbones trained using manifold mixup (Verma et al., 2019). Since these models are explicitly trained to generalize across interpolated samples, we hypothesized that synthetic mixup-based OOD prototypes might no longer serve as an effective deviation reference. To test this, we evaluated GROOD on a ResNet-18 trained with manifold mixup. Instead of using mixup-based OOD prototypes since it will no longer represent OOD data, we used a mean prototype computed from ID class prototypes. The results are shown below:

These results show that GROOD maintains competitive performance by adjusting its prototype strategy to the model's training procedure. The mixup-based prototype remains optimal for standard-trained models, while alternative strategies like mean prototypes are preferable when the backbone is mixup-regularized.