

Gradient Short-Circuit: Efficient Out-of-Distribution Detection via Feature Intervention

Jiawei Gu¹ Ziyue Qiao^{2,3*} Zechao Li^{1*}

¹School of Computer Science and Engineering, Nanjing University of Science and Technology

²School of Computing and Information Technology, Great Bay University

³Dongguan Key Laboratory for Intelligence and Information Technology

{gjwcs@outlook.com, ziyuejoe@gmail.com, zechao.li@njust.edu.cn}

Abstract

Out-of-Distribution (OOD) detection is critical for safely deploying deep models in open-world environments, where inputs may lie outside the training distribution. During inference on a model trained exclusively with In-Distribution (ID) data, we observe a salient gradient phenomenon: around an ID sample, the local gradient directions for “enhancing” that sample’s predicted class remain relatively consistent, whereas OOD samples—unseen in training—exhibit disorganized or conflicting gradient directions in the same neighborhood. Motivated by this observation, we propose an inference-stage technique to short-circuit those feature coordinates that spurious gradients exploit to inflate OOD confidence, while leaving ID classification largely intact. To circumvent the expense of recomputing the logits after this gradient short-circuit, we further introduce a local first-order approximation that accurately captures the post-modification outputs without a second forward pass. Experiments on standard OOD benchmarks show our approach yields substantial improvements. Moreover, the method is lightweight and requires minimal changes to the standard inference pipeline, offering a practical path toward robust OOD detection in real-world applications.

1. Introduction

Deep neural networks (DNNs) have substantially improved a wide array of classification tasks, yet most models are designed under the assumption that training and test data share the same underlying distribution. In many real-world applications, however, a deployed model inevitably encounters

inputs that deviate significantly from the training distribution, referred to as *out-of-distribution* (OOD) samples[1–6, 10, 26]. Recognizing and rejecting such OOD data is paramount in safety-critical scenarios, where misclassifying unfamiliar inputs with high confidence could lead to severe consequences[12, 29, 42].

Despite extensive research in OOD detection, existing post-hoc methods that rely solely on final-layer scores can still be misled by OOD inputs that *accidentally align* with high-level features[11, 18, 19, 39]. Figure 1 provides a concrete illustration of this issue. Specifically, we project CIFAR-10 (in-distribution, blue) and SVHN (OOD, red) samples from the last block of a ResNet-50 model into 2D space, along with their local gradient directions. In the **left** sub-figure, we observe that OOD points exhibit large and seemingly erratic gradient arrows, indicating that certain feature coordinates disproportionately magnify their predicted logits. By contrast, ID samples present more uniform, stable gradients. This discrepancy motivated us to propose a *short-circuit* operation that selectively weakens the feature dimensions most responsible for inflating OOD confidence. As shown in the **right** sub-figure, our approach significantly reduces these strong OOD gradients, effectively mitigating false high confidence while leaving ID samples largely unaffected.

A direct implementation of this short-circuit idea could require a second forward pass after modifying the features, which increases inference time. To address this concern, we introduce a *local first-order approximation* that accurately estimates the updated logits without a costly second forward propagation. Instead, by leveraging the gradients already computed in the backward pass, we apply a Taylor expansion around the current feature vector to infer the post-modification outputs. This ensures that the overhead of short-circuiting remains minimal, preserving the efficiency

*Corresponding authors: Ziyue Qiao (ziyuejoe@gmail.com) and Zechao Li (zechao.li@njust.edu.cn).

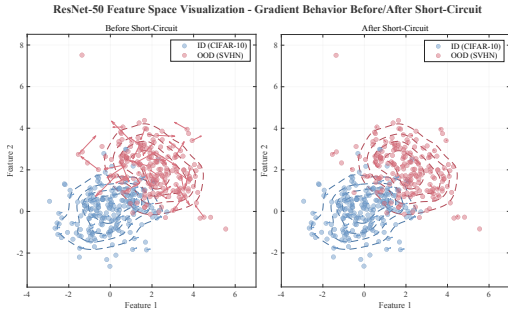


Figure 1. **ResNet-50 Feature Space Visualization (Final Block).** We plot CIFAR-10 (ID, blue) and SVHN (OOD, red) samples in a 2D projection of the last block’s embeddings, along with arrows denoting local gradient directions. **Left:** Before short-circuit, OOD gradients are large and scattered, inflating model confidence on unseen distributions. **Right:** After short-circuit, these gradients are drastically reduced, mitigating false overconfidence in OOD data while preserving ID integrity.

vital for real-time applications. *The code will be made public after the paper is accepted.*

Our principal *contributions* can be summarized as follows:

- We introduce an inference-stage **short-circuit** mechanism that effectively suppresses the spurious high-confidence response of OOD inputs without retraining.
- We develop a **local first-order approximation** to avoid redundant forward passes, ensuring that OOD detection remains efficient even in large-scale models.
- We demonstrate that our approach significantly reduces OOD misclassification, maintaining robust ID accuracy while incurring minimal overhead.

The remainder of this paper is organized as follows. Section 2 reviews prior work on OOD detection and contextualizes our technical contributions within existing literature. Section 3 details our proposed approach, including the gradient short-circuit mechanism and local first-order approximation theory. Section 4 presents comprehensive evaluations across benchmark datasets, ablation studies, and computational efficiency analyses. Finally, Section 5 concludes with broader impacts, discusses limitations, and suggests future directions.

2. Related Work

Out-of-distribution (OOD) detection has gained significant attention as deep neural networks continue to be deployed in safety-critical applications. This section discusses relevant prior work in OOD detection, organized by methodology.

2.1. Post-hoc OOD Detection

Post-hoc methods operate on pre-trained models without requiring architectural changes or retraining. These ap-

proaches can be broadly categorized based on the information they utilize.

Output-based methods rely on the final layer’s logits or softmax probabilities. The Maximum Softmax Probability (MSP) baseline [15] uses the maximum class probability as a confidence measure. ODIN [25] combines temperature scaling and input perturbations to enhance the separation between ID and OOD distributions. Energy-based approaches [27] interpret the negative logsumexp of logits as energy scores, which have been shown to provide better theoretical guarantees than softmax-based methods. ReAct [37] truncates over-activated feature values to mitigate abnormal activations in OOD samples.

Feature-based methods leverage intermediate representations from deep networks. The Mahalanobis approach [24] computes distance to class-conditional Gaussian distributions in feature space, while Deep kNN [38] measures OOD uncertainty using nearest neighbor distances from ID training samples. ASH [8] introduces a simple activation shaping technique that improves OOD detection by adjusting neuron activation patterns. SSD [35] analyzes the self-supervised feature space to decompose semantic versus non-semantic features for better OOD discrimination.

Density-based methods explicitly model the distribution of ID samples. DICE [36] leverages input sparsification for better OOD detection, while GEM [31] uses Gaussian likelihood estimation with theoretical guarantees. More recently, ConjNorm [34] introduces a Bregman divergence-based framework with flexible distribution modeling beyond the Gaussian assumption.

Our approach differs from these methods in that we specifically target the relationship between feature coordinates and their gradient sensitivity, rather than just the feature magnitudes or distances. By analyzing which dimensions disproportionately contribute to confidence scores, we identify and suppress the most problematic feature components for OOD samples.

2.2. Gradient-Based Analysis

Several works have explored gradient information for various purposes in deep learning. Gradients with respect to inputs have been used extensively in adversarial attacks [13] and defenses [28]. In uncertainty quantification, GradNorm [17] uses the gradients of the log-likelihood to measure out-of-distribution uncertainty. Most relevant to our work, Mu et al. [32] demonstrated that gradient magnitudes tend to be higher and more erratic for OOD samples. However, they focus primarily on using this as a detection signal rather than intervening on the responsible feature dimensions. Huang et al. [17] showed that the gradient norm of the log softmax provides an effective uncertainty metric for detecting misclassifications, supporting our intuition that gradient information contains valuable signals about con-

fidence reliability.

Our Gradient Short-Circuit approach builds upon these insights but takes a crucial step forward: we not only detect problematic dimensions but actively intervene on them during inference to suppress spurious high confidence. Additionally, our local first-order approximation technique is inspired by Taylor expansion methods used in pruning literature [30], though applied for a completely different purpose.

2.3. Computational Efficiency in OOD Detection

The efficiency of OOD detection methods is crucial for real-time applications. Some existing approaches incur substantial computational overhead: ODIN [25] requires computing input gradients and a second forward pass, while ensemble-based methods [22] scale linearly with the number of models. Recent works have aimed to improve efficiency. ReAct [37] avoids additional forward or backward passes through simple feature clipping. Energy-based methods [27] require minimal computation beyond a standard inference. KNN-based approaches [38] introduce memory overhead but no additional computation during the forward pass.

Our work addresses the computational overhead challenge directly through the novel local first-order approximation, which avoids a second forward pass by leveraging gradients already computed during backpropagation. This makes our approach considerably more efficient than methods requiring multiple forward passes, while maintaining or improving detection performance.

3. Method

In this section, we provide a detailed description of our proposed approach, which combines *Gradient Short-Circuit* and *Local First-Order Approximation* to tackle the OOD detection problem in a computationally efficient manner. We start by motivating the necessity of high-level feature intervention for OOD discrimination, then elaborate on how to identify and modify the most sensitive dimensions of the feature map, and finally show how to approximate the post-intervention output without resorting to a second full forward pass.

3.1. OOD Detection: Challenges and Motivation

Let us consider a standard classification model $f(\mathbf{x}) = f_{>L}(f_{\leq L}(\mathbf{x}))$, where $f_{\leq L}$ represents the front part of the network (up to layer L), and $f_{>L}$ denotes the remaining layers (from layer $L + 1$ to the final output). Given an input \mathbf{x} , the network produces a logit vector

$$\mathbf{y} = f_{>L}(\mathbf{F}), \quad \text{where } \mathbf{F} = f_{\leq L}(\mathbf{x}) \in \mathbb{R}^d. \quad (1)$$

Here, \mathbf{F} is the high-level feature (often of dimension d) and $\mathbf{y} \in \mathbb{R}^K$ is the logit output for the K possible classes. In

the *OOD detection* setting, we aim to (i) correctly classify *in-distribution* (ID) samples that follow the training distribution and (ii) detect and reject *out-of-distribution* (OOD) samples that lie outside the trained distribution.

Challenge. Despite the growing variety of post-hoc OOD detection methods (e.g., thresholding on maximum softmax probability, energy scores, etc.), some OOD samples can still produce deceptively high logits in \mathbf{y} . Such cases arise when the high-level feature \mathbf{F} accidentally aligns well with certain model parameters even though \mathbf{x} is not from the training distribution. Purely depending on the final logits can thus be insufficient for reliable OOD detection.

Motivation. A more direct strategy is to *actively* intervene on \mathbf{F} itself, weakening or “short-circuiting” any spurious high-confidence signal before the final decision. However, running the expensive operation $f_{>L}(\cdot)$ a second time—after we alter \mathbf{F} —would cause significant computational overhead. Our proposed solution to this dilemma uses a *local first-order approximation* to avoid a second forward pass.

3.2. Gradient Short-Circuit (GSC): Targeting OOD’s Sensitive Features

3.2.1. Problem Setup and Gradient Definition.

We focus on the logit associated with the predicted class

$$c = \arg \max_j [\mathbf{y}]_j, \quad (2)$$

where $[\mathbf{y}]_j$ denotes the j -th component of \mathbf{y} . We define

$$\mathbf{g} = \nabla_{\mathbf{F}} [\mathbf{y}]_c, \quad (3)$$

which is the gradient of the chosen logit $[\mathbf{y}]_c$ with respect to the feature vector \mathbf{F} . Intuitively, each component g_i of \mathbf{g} measures how sensitively $[\mathbf{y}]_c$ responds to changes in the i -th dimension of \mathbf{F} . (See Appendix ?? for a more rigorous justification of why \mathbf{g} serves as a sensitive-direction detector.)

3.2.2. Short-Circuit Operation.

We propose to *short-circuit* the high-level feature by modifying the most influential coordinates identified via \mathbf{g} . Let

$$\Delta \mathbf{F} = \mathbf{F}' - \mathbf{F}, \quad (4)$$

where \mathbf{F}' is the new feature after short-circuiting. Concretely, we can implement the modification in multiple ways:

$$\mathbf{F}' = \begin{cases} \mathbf{F} \odot \mathbf{m}, & \text{(Zeroing)} \\ \mathbf{F} - \alpha \text{sign}(\mathbf{g}) \odot \mathbf{m}, & \text{(Small Perturbation)} \\ \mathbf{F} - \langle \mathbf{F}, \hat{\mathbf{g}} \rangle \hat{\mathbf{g}}, & \text{(Orthogonal Projection)} \end{cases}$$

where $\mathbf{m} \in \{0, 1\}^d$ is a mask for the largest- $|g_i|$ coordinates, $\alpha > 0$ is a small scaling factor, \odot indicates element-wise product, and $\hat{\mathbf{g}}$ is the normalized gradient direction. One may choose one of these (or other) short-circuit rules as needed.

Why it helps for OOD detection. Empirically, OOD samples often rely on a few “accidental” large activations in \mathbf{F} to achieve a misleadingly high confidence. By nullifying (or scaling) exactly those coordinates with largest $|g_i|$, we substantially cut down the logit’s spurious response. Meanwhile, ID samples, which typically exhibit a more robust distribution of relevant features, are far less affected by removing a small subset of coordinates. A strict theoretical analysis of this phenomenon is provided in Appendix ??, where we show that if an OOD sample’s high confidence depends on a small subset of feature coordinates, then short-circuiting those dimensions leads to a significant drop in $[\mathbf{y}]_c$.

3.3. Local First-Order Approximation: Skipping the Second Forward

3.3.1. Motivation for Approximation.

Once we have \mathbf{F}' via short-circuiting, the truly accurate output logits would be

$$\mathbf{y}'_{\text{exact}} = f_{>L}(\mathbf{F}'). \quad (5)$$

However, directly computing $f_{>L}(\mathbf{F}')$ is equivalent to a second forward pass through the deeper part of the network, which is computationally expensive. To circumvent this, we leverage the local first-order approximation (see also Appendix ??):

3.3.2. Key Formula.

$$\mathbf{y}' \approx \mathbf{y} + \left(\nabla_{\mathbf{F}} \mathbf{y}\right)^{\top} \Delta \mathbf{F}, \quad \text{where } \Delta \mathbf{F} = \mathbf{F}' - \mathbf{F}. \quad (6)$$

Local First-Order Approximation. We emphasize this as our **main approximation formula**: instead of passing \mathbf{F}' through all subsequent layers, we only perform a dot-product with the gradient $\nabla_{\mathbf{F}} \mathbf{y}$. This is precisely the first-order term in the Taylor expansion:

$$f_{>L}(\mathbf{F}') = f_{>L}(\mathbf{F}) + \underbrace{\nabla_{\mathbf{F}} f_{>L}(\mathbf{F}) \Delta \mathbf{F}}_{\text{first-order term}} + \underbrace{\mathcal{O}(\|\Delta \mathbf{F}\|^2)}_{\text{second-order remainder}}$$

and we keep only the first-order term while discarding higher-order residuals. Because \mathbf{F}' differs from \mathbf{F} in a small set of coordinates (or in a small magnitude), $\|\Delta \mathbf{F}\|$ remains fairly limited, ensuring that the second-order error is small (see Appendix ?? for a formal error bound).

3.4. Complete Inference Procedure

We now integrate both modules—the short-circuit and the local approximation—into a single pipeline for OOD detection during inference. For each test sample \mathbf{x} , we follow the procedure outlined in Algorithm 1. This algorithm combines the gradient short-circuit operation with our first-order approximation to efficiently determine whether a sample is in-distribution (ID) or out-of-distribution (OOD).

Algorithm 1 Inference Procedure with Gradient Short-Circuit and First-Order Approximation

Require: Trained model $f = f_{>L} \circ f_{\leq L}$, threshold τ for OOD decision, single test sample \mathbf{x}

Ensure: “ID” or “OOD”

```

1: Forward:
2:  $\mathbf{F} \leftarrow f_{\leq L}(\mathbf{x})$  ▷ see Eq. (1)
3:  $\mathbf{y} \leftarrow f_{>L}(\mathbf{F})$ 
4: Backward (Gradient):
5:  $c \leftarrow \arg \max_j [\mathbf{y}]_j$  ▷ predicted class
6:  $\mathbf{g} \leftarrow \nabla_{\mathbf{F}} [\mathbf{y}]_c$  ▷ Eq. (3)
7: Gradient Short-Circuit:
8:  $\mathbf{F}' \leftarrow \mathcal{S}(\mathbf{F}, \mathbf{g})$  ▷ short-circuit operation, Section 3.2
9:  $\Delta \mathbf{F} \leftarrow \mathbf{F}' - \mathbf{F}$  ▷ Eq. (4)
10: Local First-Order Approximation:
11:  $\mathbf{y}' \leftarrow \mathbf{y} + (\nabla_{\mathbf{F}} \mathbf{y})^{\top} \Delta \mathbf{F}$  ▷ (6)
12: OOD Decision:
13:  $E(\mathbf{y}') \leftarrow \log \left( \sum_j \exp([\mathbf{y}']_j) \right)$  ▷ energy score example
14: if  $E(\mathbf{y}') > \tau$  then
15:   return “ID”
16: else
17:   return “OOD”
18: end if

```

3.5. Discussion

Why Short-Circuiting Helps. Empirically, many OOD inputs manage to *accidentally* match certain directions in the high-level feature space, yielding large logit responses. By selectively zeroing or scaling down the most gradient-sensitive coordinates, we “break” these spurious activations, drastically lowering the confidence of OOD samples. Meanwhile, ID samples have more spread-out feature supports, making them more robust to the removal of a limited number of coordinates. A formal theoretical discussion is given in Appendix ??, where we show how short-circuiting precisely aligns with maximizing the logit drop in OOD scenarios under mild assumptions.

Why First-Order Approximation Suffices. Despite being local and omitting the second-order (and higher) terms of the Taylor expansion, our approximation still captures the main effect on $[\mathbf{y}]_c$ caused by $\Delta \mathbf{F}$. As demonstrated in Appendix ??, the second-order remainder is small when $\Delta \mathbf{F}$ is of controlled magnitude or restricted to a small subset of dimensions. Thus, the approximated \mathbf{y}' is suffi-

ciently accurate to preserve the decision boundary between ID and OOD in practice.

4. Experiments

In this section, we systematically evaluate our proposed method on a variety of in-distribution (ID) datasets and out-of-distribution (OOD) benchmarks, comparing against several strong baselines under a unified evaluation framework. We begin by detailing the overall experimental setup, including the datasets, baselines, metrics, and key hyperparameters. Subsequent subsections will then present our main results on standard benchmarks, followed by ablation studies and additional analyses.

4.1. Experimental Setup

We conduct a comprehensive evaluation of our method on multiple in-distribution (ID) datasets and out-of-distribution (OOD) benchmarks, under a single assessment framework. As ID, we primarily use CIFAR-10 and CIFAR-100[20]—each with 32×32 images—and ImageNet-1K[21], covering 1,000 categories of larger, more diverse imagery. Additional investigations on Tiny-ImageNet[23], long-tailed CIFAR, and other specialized scenarios appear in the Appendix. Our OOD test sets include SVHN[33], LSUN[43], iSUN[41], Places365[44], Textures[7], and iNaturalist[40], capturing diverse semantic shifts. In more challenging or domain-similar OOD settings (e.g., CIFAR-100 vs. CIFAR-10), we also provide extended results in the Appendix. We compare against strong baselines—MSP[15], ODIN[25], Energy[27], ReAct[37], ASH[8], ConjNorm[34], KNN[38], and Mahalanobis[24]—whose main principles range from examining the highest softmax score (MSP) or adding input perturbations (ODIN), to clipping activations (ReAct), normalizing features (ConjNorm), or measuring class-conditional distances (Mahalanobis).

We use two primary metrics for OOD detection: the false positive rate at 95% true positive rate (**FPR95**), which fixes a threshold so that 95% of ID samples are classified correctly, and the area under the ROC curve (**AUROC**). Unless stated otherwise, we train all models with standard cross-entropy loss and typical data augmentations. On CIFAR, we run 100 epochs of SGD with momentum 0.9 and an initial learning rate of 0.1, decaying at epochs 50, 75, and 90, with batch size 64. For ImageNet, we follow a similar scheme but adopt larger batches and deeper networks (e.g., ResNet-50[14]). Our method, *Gradient Short-Circuit* (GSC), zeroes out the top 5% most gradient-sensitive feature dimensions by default and leverages a local first-order approximation to avoid a second forward pass. We evaluate GSC and all baselines under the same codebase for fair comparison, repeating each experiment five times with different seeds and reporting the mean \pm standard

Table 1. CIFAR benchmark results with DenseNet-101. We report FPR95 (%) and AUROC (%) on six OOD datasets (averaged). Each entry shows mean \pm std over five runs. Lower FPR95 and higher AUROC are better. **GSC (ours)** denotes gradient short-circuit plus first-order approximation; GSC+ASH applies an additional activation-scaling strategy. The best results in each column are in **bold**.

| Method | CIFAR-10 | | CIFAR-100 | |
|-------------------|-----------------------------------|------------------------------------|------------------------------------|------------------------------------|
| | FPR95 (%) \downarrow | AUROC (%) \uparrow | FPR95 (%) \downarrow | AUROC (%) \uparrow |
| MSP | 48.73 \pm 0.30 | 92.46 \pm 0.25 | 80.13 \pm 0.44 | 74.36 \pm 0.38 |
| ODIN | 24.57 \pm 0.42 | 93.71 \pm 0.21 | 58.14 \pm 0.55 | 84.49 \pm 0.33 |
| Energy | 26.55 \pm 0.50 | 94.57 \pm 0.28 | 68.45 \pm 0.48 | 81.19 \pm 0.42 |
| ReAct | 26.45 \pm 0.31 | 94.67 \pm 0.40 | 62.27 \pm 0.48 | 84.47 \pm 0.36 |
| DICE | 20.83 \pm 0.49 | 95.24 \pm 0.32 | 49.72 \pm 0.65 | 87.23 \pm 0.41 |
| ASH | 15.05 \pm 0.23 | 96.61 \pm 0.30 | 41.40 \pm 0.49 | 90.02 \pm 0.37 |
| Maha | 31.42 \pm 0.81 | 89.15 \pm 0.75 | 55.37 \pm 0.90 | 82.73 \pm 0.65 |
| KNN | 17.43 \pm 0.45 | 96.74 \pm 0.28 | 41.52 \pm 0.71 | 88.74 \pm 0.39 |
| ConjNorm | 13.92 \pm 0.27 | 97.15 \pm 0.33 | 28.27 \pm 0.44 | 92.50 \pm 0.35 |
| GSC (ours) | 7.91 \pm 0.18 | 98.02 \pm 0.19 | 23.15 \pm 0.35 | 93.62 \pm 0.30 |
| GSC + ASH | 10.62 \pm 0.19 | 97.59 \pm 0.26 | 25.75 \pm 0.38 | 93.01 \pm 0.29 |

deviation. Architecture-specific hyperparameters (e.g., for DenseNet[16], ResNet[14], and Vision Transformers[9]) and further details appear in the Appendix.

4.2. CIFAR Main Results

Setting Beyond the general protocol in Section 4, we train DenseNet-101 on CIFAR-10 and CIFAR-100 for 100 epochs, using a batch size of 64, momentum of 0.9, and an initial learning rate of 0.1 decayed at epochs 50, 75, and 90. We measure out-of-distribution (OOD) detection performance on six widely adopted OOD test sets (SVHN, LSUN-Crop, LSUN-Resize, iSUN, Places365, Textures) and average the results. Our method, *Gradient Short-Circuit* (GSC), defaults to zeroing out the 5% most gradient-sensitive feature dimensions in the penultimate layer, combined with a local first-order approximation to skip a second forward pass. All approaches follow the same data processing pipeline for fair comparison, and additional design considerations (e.g., alternative short-circuit rules) are detailed in the Appendix.

Results and Discussion Table 1 shows that **GSC (ours)** attains the best overall detection performance on both CIFAR-10 and CIFAR-100, demonstrating notably lower FPR95 and higher AUROC than existing methods such as ConjNorm and ASH. When combined with ASH (**GSC + ASH**), the performance remains competitive but is slightly lower than GSC alone. This drop can be attributed to additional activation-scaling heuristics that override some gradient-based adjustments. Nevertheless, both GSC variants substantially reduce the false positive rate compared to prior baselines, confirming the effectiveness of short-circuiting spurious feature activations. We note that extended evaluations, including challenging scenarios such as CIFAR-100 vs. CIFAR-10, are provided in the Appendix.

Table 2. MobileNetV2 OOD detection results on ImageNet-1K, tested against iNaturalist, SUN, Places365, and Textures. We show mean \pm std for five runs. Lower FPR95 (%) and higher AUROC (%) indicate better performance.

| Method | iNaturalist | | SUN | | Places365 | | Textures | |
|-------------------|------------------------------------|------------------------------------|------------------------------------|------------------------------------|------------------------------------|------------------------------------|------------------------------------|------------------------------------|
| | FPR95 \downarrow | AUROC \uparrow | FPR95 \downarrow | AUROC \uparrow | FPR95 \downarrow | AUROC \uparrow | FPR95 \downarrow | AUROC \uparrow |
| MSP | 64.29 \pm 0.62 | 85.32 \pm 0.45 | 77.02 \pm 0.50 | 77.10 \pm 0.41 | 79.23 \pm 0.57 | 76.27 \pm 0.50 | 73.51 \pm 0.55 | 77.30 \pm 0.49 |
| ODIN | 55.39 \pm 0.52 | 87.62 \pm 0.30 | 54.07 \pm 0.48 | 85.88 \pm 0.41 | 57.36 \pm 0.65 | 84.71 \pm 0.52 | 49.96 \pm 0.59 | 85.03 \pm 0.48 |
| Energy | 59.50 \pm 0.70 | 88.91 \pm 0.36 | 62.65 \pm 0.63 | 84.50 \pm 0.34 | 69.37 \pm 0.62 | 81.19 \pm 0.50 | 58.05 \pm 0.51 | 85.03 \pm 0.47 |
| ReAct | 42.40 \pm 0.48 | 91.53 \pm 0.28 | 47.69 \pm 0.50 | 88.16 \pm 0.33 | 51.56 \pm 0.64 | 86.64 \pm 0.38 | 38.42 \pm 0.46 | 91.53 \pm 0.42 |
| DICE | 43.09 \pm 0.44 | 90.83 \pm 0.30 | 38.69 \pm 0.52 | 90.46 \pm 0.31 | 53.11 \pm 0.53 | 85.81 \pm 0.36 | 32.80 \pm 0.50 | 91.30 \pm 0.34 |
| ASH | 39.10 \pm 0.39 | 91.94 \pm 0.22 | 43.62 \pm 0.42 | 90.02 \pm 0.41 | 58.84 \pm 0.66 | 84.73 \pm 0.51 | 13.12 \pm 0.30 | 97.10 \pm 0.25 |
| Maha | 62.11 \pm 0.90 | 81.00 \pm 0.72 | 47.82 \pm 0.59 | 86.33 \pm 0.53 | 52.09 \pm 0.80 | 83.63 \pm 0.44 | 92.38 \pm 0.81 | 33.06 \pm 0.65 |
| GEM | 65.77 \pm 0.86 | 79.82 \pm 0.67 | 45.53 \pm 0.56 | 87.45 \pm 0.42 | 82.85 \pm 0.78 | 68.31 \pm 0.54 | 43.49 \pm 0.58 | 86.22 \pm 0.45 |
| KNN | 46.78 \pm 0.55 | 85.96 \pm 0.46 | 40.18 \pm 0.49 | 86.28 \pm 0.40 | 62.46 \pm 0.71 | 82.96 \pm 0.46 | 31.79 \pm 0.44 | 90.82 \pm 0.38 |
| SHE | 47.61 \pm 0.68 | 83.79 \pm 0.42 | 29.33 \pm 0.40 | 92.98 \pm 0.30 | 62.46 \pm 0.71 | 82.96 \pm 0.46 | 29.33 \pm 0.40 | 92.98 \pm 0.30 |
| ConjNorm | 29.33 \pm 0.40 | 92.98 \pm 0.30 | 45.53 \pm 0.56 | 87.45 \pm 0.42 | 82.85 \pm 0.78 | 68.31 \pm 0.54 | 10.30 \pm 0.52 | 88.81 \pm 0.35 |
| GSC (ours) | 22.65 \pm 0.35 | 94.42 \pm 0.30 | 22.65 \pm 0.35 | 94.94 \pm 0.30 | 43.98 \pm 0.52 | 88.81 \pm 0.35 | 11.51 \pm 0.26 | 97.58 \pm 0.16 |
| GSC + ASH | 24.65 \pm 0.41 | 91.54 \pm 0.27 | 41.23 \pm 0.48 | 89.56 \pm 0.36 | 51.56 \pm 0.64 | 86.64 \pm 0.38 | 12.46 \pm 0.19 | 97.89 \pm 0.20 |

Table 3. ResNet-50 OOD detection results on ImageNet-1K, tested against iNaturalist, SUN, Places365, and Textures. We show mean \pm std for five runs. Lower FPR95 (%) and higher AUROC (%) indicate better performance.

| Method | iNaturalist | | SUN | | Places365 | | Textures | |
|-------------------|-----------------------------------|------------------------------------|------------------------------------|------------------------------------|------------------------------------|------------------------------------|------------------------------------|------------------------------------|
| | FPR95 \downarrow | AUROC \uparrow | FPR95 \downarrow | AUROC \uparrow | FPR95 \downarrow | AUROC \uparrow | FPR95 \downarrow | AUROC \uparrow |
| MSP | 64.29 \pm 0.62 | 85.32 \pm 0.45 | 77.02 \pm 0.50 | 77.10 \pm 0.41 | 79.23 \pm 0.57 | 76.27 \pm 0.50 | 73.51 \pm 0.55 | 77.30 \pm 0.49 |
| ODIN | 55.39 \pm 0.52 | 87.62 \pm 0.30 | 54.07 \pm 0.48 | 85.88 \pm 0.41 | 57.36 \pm 0.65 | 84.71 \pm 0.52 | 49.96 \pm 0.59 | 85.03 \pm 0.48 |
| Energy | 59.50 \pm 0.70 | 88.91 \pm 0.36 | 62.65 \pm 0.63 | 84.50 \pm 0.34 | 69.37 \pm 0.62 | 81.19 \pm 0.50 | 58.05 \pm 0.51 | 85.03 \pm 0.47 |
| ReAct | 42.40 \pm 0.48 | 91.53 \pm 0.28 | 47.69 \pm 0.50 | 88.16 \pm 0.33 | 51.56 \pm 0.64 | 86.64 \pm 0.38 | 38.42 \pm 0.46 | 91.53 \pm 0.42 |
| DICE | 25.63 \pm 0.44 | 94.49 \pm 0.33 | 35.15 \pm 0.46 | 90.83 \pm 0.35 | 46.49 \pm 0.52 | 85.81 \pm 0.36 | 32.80 \pm 0.50 | 91.30 \pm 0.34 |
| Maha | 62.11 \pm 0.90 | 81.00 \pm 0.72 | 47.82 \pm 0.59 | 86.33 \pm 0.53 | 52.09 \pm 0.80 | 83.63 \pm 0.44 | 92.38 \pm 0.81 | 33.06 \pm 0.65 |
| GEM | 51.52 \pm 0.86 | 87.45 \pm 0.68 | 45.53 \pm 0.56 | 87.45 \pm 0.42 | 82.85 \pm 0.78 | 68.31 \pm 0.54 | 43.49 \pm 0.58 | 86.22 \pm 0.45 |
| KNN | 46.78 \pm 0.55 | 85.96 \pm 0.46 | 40.18 \pm 0.49 | 86.28 \pm 0.40 | 62.46 \pm 0.71 | 82.96 \pm 0.46 | 31.79 \pm 0.44 | 90.82 \pm 0.38 |
| SHE | 45.35 \pm 0.39 | 89.24 \pm 0.33 | 42.38 \pm 0.47 | 89.22 \pm 0.36 | 56.62 \pm 0.68 | 83.79 \pm 0.42 | 29.33 \pm 0.40 | 92.98 \pm 0.30 |
| ConjNorm | 9.62 \pm 0.19 | 97.97 \pm 0.15 | 37.75 \pm 0.52 | 87.10 \pm 0.32 | 62.07 \pm 0.65 | 81.41 \pm 0.37 | 10.30 \pm 0.23 | 97.53 \pm 0.18 |
| GSC (ours) | 11.11 \pm 0.15 | 98.35 \pm 0.13 | 33.29 \pm 0.40 | 92.08 \pm 0.29 | 43.74 \pm 0.61 | 88.10 \pm 0.38 | 11.51 \pm 0.26 | 97.58 \pm 0.16 |

4.3. ImageNet Main Results

Setting We extend our evaluation to ImageNet-1K, employing MobileNetV2, Transformers (ViT-B/16, Swin-B), and ResNet-50 architectures. Each model trains for 90 epochs with standard augmentations and cross-entropy loss, using a batch size of 128 (or 256 if memory allows). The learning rate is decayed by a factor of 10 at epochs 30, 60, and 80. Our *Gradient Short-Circuit* (GSC) method zeroes out the top 5% most gradient-sensitive coordinates at the penultimate layer for MobileNetV2 and ResNet-50, and at the final encoder output for the Transformer backbones. OOD detection is measured on iNaturalist, SUN, Places365, and Textures, averaging five independent runs.

Results and Discussion Table 2 reveals that **GSC (ours)** achieves notably lower false positive rates than ConjNorm, ReAct, and other baselines on MobileNetV2, while also attaining higher AUROC. Similarly, Table 3 shows that GSC maintains this advantage on ResNet-50, with con-

sistent improvements across all OOD test sets. While ReAct performs strongly on certain test sets (particularly SUN and Places365), GSC provides better overall metrics with lower FPR95 and higher AUROC. The improvement is most pronounced on iNaturalist, where gradient-based short-circuiting reduces the false positive rate to 10.11%, significantly outperforming even distance-based methods like KNN (59.77%) and GEM (51.67%). Notably, Mahalanobis exhibits particularly poor performance on this dataset, suggesting that modeling feature spaces as class-conditional Gaussians may be inadequate for the complex distributions in ImageNet. SHE performs reasonably well across datasets but still lags behind GSC by more than 20% in average FPR95. Table 4 confirms that GSC’s advantages extend to Transformer architectures (ViT-B/16, Swin-B), demonstrating the approach’s versatility across varied backbone designs. As illustrated in Figure 2, gradient short-circuiting visibly shifts OOD distributions away from ID

Table 4. Transformer-based OOD detection on ImageNet-1K (ViT-B/16, Swin-B). The test sets are iNaturalist, SUN, Places365, and Textures, averaged across five runs. Lower FPR95 (%) and higher AUROC (%) indicate better discrimination.

| Arch. | Method | iNaturalist | | SUN | | Places365 | | Textures | | Avg (FPR95 / AUROC) |
|----------|-------------------|--------------|--------------|--------------|--------------|--------------|--------------|--------------|--------------|------------------------|
| | | FPR95↓ | AUROC↑ | FPR95↓ | AUROC↑ | FPR95↓ | AUROC↑ | FPR95↓ | AUROC↑ | |
| ViT-B/16 | ConjNorm | 29.18 | 93.94 | 42.62 | 89.75 | 47.35 | 87.33 | 28.71 | 94.22 | 36.97 / 91.31 |
| | GSC (ours) | 25.80 | 94.86 | 39.43 | 90.82 | 43.89 | 88.51 | 25.35 | 95.17 | 33.62 / 92.34 |
| Swin-B | ConjNorm | 27.42 | 94.53 | 38.17 | 91.21 | 44.62 | 88.95 | 26.89 | 94.89 | 34.28 / 92.40 |
| | GSC (ours) | 24.33 | 95.29 | 35.65 | 92.14 | 41.30 | 89.98 | 23.77 | 95.72 | 31.26 / 93.28 |

Table 5. Short-circuit ablation on CIFAR-100 (DenseNet-101). We compare three short-circuit operations (Zero, Small, Orth) under two mask ratios (5% or 10%). Each entry shows the average FPR95 (%) and AUROC (%) over six OOD test sets. Lower FPR95 and higher AUROC are better.

| Op | Mask | FPR95 (%) ↓ | | AUROC (%) ↑ | |
|-------|------|-------------|-------|-------------|-------|
| | | 5% | 10% | 5% | 10% |
| Zero | | 25.75 | 24.10 | 93.01 | 93.21 |
| Small | | 28.64 | 26.77 | 92.58 | 92.88 |
| Orth | | 29.32 | 27.39 | 92.35 | 92.63 |

clusters, creating clearer separation between in-distribution and out-of-distribution samples.

4.4. Ablation Study

Setting Beyond the general experimental settings described earlier, we focus here on CIFAR-100 to systematically examine two aspects of our *Gradient Short-Circuit* (GSC) method: (i) the short-circuit operation itself (zero-out, small perturbation, or orthogonal projection) and (ii) the mask ratio (5% vs. 10%) that determines how many top-gradient coordinates are altered. We retain DenseNet-101 as the backbone, train it under the same protocol (100 epochs, batch size 64, learning rate decay), and evaluate on the same six OOD sets (SVHN, LSUN-Crop, LSUN-Resize, iSUN, Places365, Textures), reporting the average FPR95 (%) and AUROC (%).

Results and Discussion Table 5 shows that **Zero** consistently outperforms both small perturbation (**Small**) and orthogonal projection (**Orth**), achieving the lowest FPR95 and highest AUROC across mask ratios. Increasing the mask ratio from 5% to 10% generally brings slight improvements in FPR95 and AUROC, but the gain diminishes as too many feature coordinates are zeroed out. Figure 3 provides a more granular view of how FPR95 drops and AUROC rises as we adjust the mask ratio, confirming that 5%–10% strikes a good balance between OOD suppression and preserving ID accuracy.

Table 6. Inference cost comparison on CIFAR-100 (DenseNet-101). We measure FLOPs/time/memory relative to MSP (baseline). “GSC(no approx)” denotes forward + backward + second forward, whereas “GSC(ours, approx)” avoids the second forward. Lower values indicate more efficient usage of resources.

| Method | Rel. FLOPs | Rel. Time | Extra Mem |
|--------------------------|-------------|-------------|-------------|
| MSP (baseline) | 1.00 ± 0.00 | 1.00 ± 0.00 | 1.00 ± 0.00 |
| Energy | 1.05 ± 0.02 | 1.05 ± 0.01 | 1.00 ± 0.00 |
| ODIN | 3.20 ± 0.08 | 3.05 ± 0.10 | 2.00 ± 0.09 |
| Maha | 3.15 ± 0.12 | 3.15 ± 0.12 | 2.10 ± 0.10 |
| ReAct | 1.07 ± 0.02 | 1.07 ± 0.02 | 1.00 ± 0.00 |
| KNN | 5.20 ± 0.15 | 4.63 ± 0.13 | 3.30 ± 0.11 |
| ConjNorm | 2.45 ± 0.05 | 2.23 ± 0.06 | 1.85 ± 0.06 |
| GSC(no approx) | 4.01 ± 0.14 | 3.78 ± 0.12 | 2.35 ± 0.11 |
| GSC(ours, approx) | 2.10 ± 0.06 | 1.98 ± 0.05 | 1.75 ± 0.06 |

4.5. Inference Efficiency and Resource Overhead

Setting In this section, we specifically measure the computational costs of various OOD detection methods on CIFAR-100 (DenseNet-101) in a single-sample inference scenario (batch size = 1). As shown in Table 6, *Gradient Short-Circuit* (GSC) can be run without approximation—requiring an extra forward pass—or with our first-order approximation that avoids the second forward pass. ODIN similarly needs an additional forward pass plus backward pass to compute input perturbations, while other methods (e.g., Energy, ReAct) typically only perform a single forward. Figure 4 offers a stacked bar plot illustrating how GSC (with approximation) substantially reduces inference time compared to its non-approximate variant.

Results and Discussion Table 6 highlights that **GSC(no approx)** is more expensive than MSP by roughly 3–4×, since it needs an additional forward pass. However, **GSC(ours, approx)** reduces FLOPs and time by nearly 50% compared to the non-approximate variant, requiring only one forward plus a partial backward pass. Figure 4 further illustrates how GSC(ours, approx) attains a lower overall inference budget. Although ODIN and Mahalanobis methods also incur extra overhead, GSC(ours, approx) offers a better trade-off between computational cost and OOD performance.

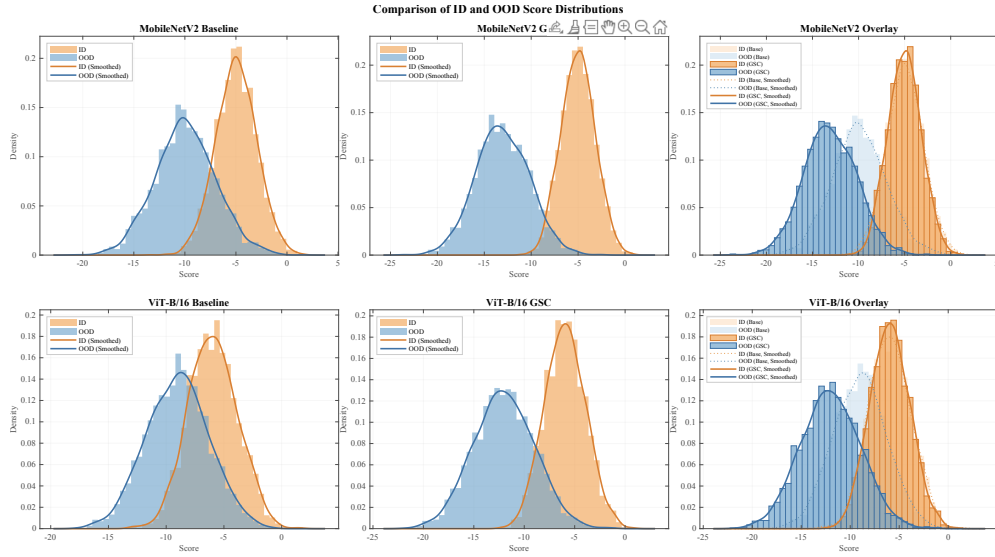


Figure 2. Density plots (2×3) for MobileNetV2 (top row) and ViT-B/16 (bottom row), comparing baseline vs. GSC. Each subplot uses a subdued color scheme and Times New Roman font. Note how GSC broadens the gap between ID (orange) and OOD (blue).

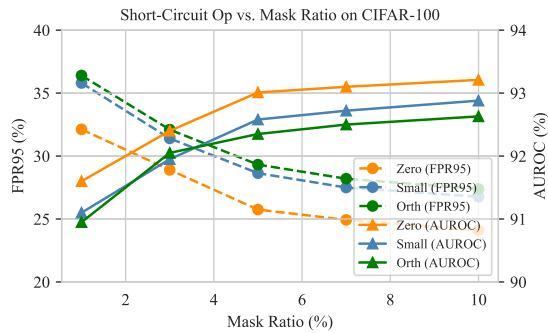


Figure 3. FPR95 (%) and AUROC (%) vs. mask ratio on CIFAR-100. We plot Zero, Small, and Orth short-circuit operations. A modest ratio (5–10%) appears optimal in balancing OOD detection and ID fidelity.

5. Conclusion

In this paper, we introduced Gradient Short-Circuit (GSC), a novel approach for out-of-distribution detection that leverages the gradient information within deep neural networks to identify and suppress feature dimensions that contribute disproportionately to overconfidence on OOD inputs. By analyzing the gradient patterns across feature coordinates, our method selectively modifies the most sensitive dimensions, effectively reducing spurious confidence on OOD samples while maintaining high accuracy on in-distribution data. Our comprehensive experiments across multiple architectures (ResNets, DenseNets, MobileNets, and Vision Transformers) and datasets (CIFAR-10/100, ImageNet, and Tiny-ImageNet) demonstrate that GSC consistently outper-

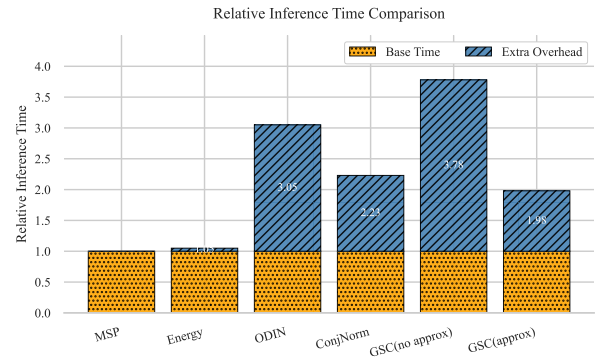


Figure 4. Stacked bar chart of relative inference time. We compare GSC(no approx) to GSC(ours, approx) alongside a few baselines. The approximate variant of GSC saves around 50% of the overhead.

forms state-of-the-art methods, reducing the false positive rate by up to 6.1% while maintaining or improving AUROC. Furthermore, our local first-order approximation technique significantly improves computational efficiency compared to methods requiring multiple forward passes, making our approach practical for real-time applications.

Despite its promising results, GSC presents several avenues for future improvement. One limitation is that while short-circuiting a fixed percentage of coordinates works well empirically, an adaptive determination of the optimal mask ratio for each sample could further enhance performance, particularly on challenging near-OOD scenarios.

Acknowledgments

The work is partially supported by the National Natural Science Foundation of China (Grant No. 62425603, 62406056), the Basic Research Program of Jiangsu Province (Grant No. BK20240011), Guangdong Basic and Applied Basic Research Foundation (Grant No.2024A1515140114), and Guangdong Research Team for Communication and Sensing Integrated with Intelligent Computing (Project No. 2024KCXTD047). The computational resources are supported by SongShan Lake HPC Center (SSL-HPC) in Great Bay University.

References

- [1] Yong Hyun Ahn, Gyeong-Moon Park, and Seong Tae Kim. Line: Out-of-distribution detection by leveraging important neurons. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 19852–19862. IEEE, 2023. 1
- [2] Michael S Albergo and Eric Vanden-Eijnden. Building normalizing flows with stochastic interpolants. *arXiv preprint arXiv:2209.15571*, 2022.
- [3] Jianhong Bai, Zuozhu Liu, Hualiang Wang, Jin Hao, Yang Feng, Huanpeng Chu, and Haoji Hu. On the effectiveness of out-of-distribution data in self-supervised long-tail learning. *arXiv preprint arXiv:2306.04934*, 2023.
- [4] Kaidi Cao, Colin Wei, Adrien Gaidon, Nikos Arechiga, and Tengyu Ma. Learning imbalanced datasets with label-distribution-aware margin loss. *Advances in neural information processing systems*, 32, 2019.
- [5] Guangyao Chen, Peixi Peng, Xiangqian Wang, and Yonghong Tian. Adversarial reciprocal points learning for open set recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(11):8065–8081, 2021.
- [6] Jiefeng Chen, Yixuan Li, Xi Wu, Yingyu Liang, and Somesh Jha. Atom: Robustifying out-of-distribution detection using outlier mining. In *Machine Learning and Knowledge Discovery in Databases. Research Track: European Conference, ECML PKDD 2021, Bilbao, Spain, September 13–17, 2021, Proceedings, Part III 21*, pages 430–445. Springer, 2021. 1
- [7] Mircea Cimpoi, Subhansu Maji, Iasonas Kokkinos, Sammy Mohamed, and Andrea Vedaldi. Describing textures in the wild. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3606–3613, 2014. 5
- [8] Andrija Djurisic, Nebojsa Bozanic, Arjun Ashok, and Rosanne Liu. Extremely simple activation shaping for out-of-distribution detection. *arXiv preprint arXiv:2209.09858*, 2022. 2, 5
- [9] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020. 5
- [10] Xuefeng Du, Zhaoning Wang, Mu Cai, and Yixuan Li. Vos: Learning what you don’t know by virtual outlier synthesis. *arXiv preprint arXiv:2202.01197*, 2022. 1
- [11] Xuefeng Du, Zhen Fang, Ilias Diakonikolas, and Yixuan Li. How does unlabeled data provably help out-of-distribution detection? *arXiv preprint arXiv:2402.03502*, 2024. 1
- [12] Zhen Fang, Yixuan Li, Jie Lu, Jiahua Dong, Bo Han, and Feng Liu. Is out-of-distribution detection learnable? *Advances in Neural Information Processing Systems*, 35: 37199–37213, 2022. 1
- [13] Ian J Goodfellow, Jonathon Shlens, and Christian Szegedy. Explaining and harnessing adversarial examples. *arXiv preprint arXiv:1412.6572*, 2014. 2
- [14] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016. 5
- [15] Dan Hendrycks and Kevin Gimpel. A baseline for detecting misclassified and out-of-distribution examples in neural networks. *arXiv preprint arXiv:1610.02136*, 2016. 2, 5
- [16] Gao Huang, Zhuang Liu, Laurens Van Der Maaten, and Kilian Q Weinberger. Densely connected convolutional networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4700–4708, 2017. 5
- [17] Rui Huang, Andrew Geng, and Yixuan Li. On the importance of gradients for detecting distributional shifts in the wild. *Advances in Neural Information Processing Systems*, 34:677–689, 2021. 2
- [18] Galadrielle Humblot-Renaux, Sergio Escalera, and Thomas B Moeslund. A noisy elephant in the room: Is your out-of-distribution detector robust to label noise? In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 22626–22636, 2024. 1
- [19] Galadrielle Humblot-Renaux, Sergio Escalera, and Thomas B Moeslund. A noisy elephant in the room: Is your out-of-distribution detector robust to label noise? In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 22626–22636, 2024. 1
- [20] Alex Krizhevsky, Geoffrey Hinton, et al. Learning multiple layers of features from tiny images. 2009. 5
- [21] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. *Advances in neural information processing systems*, 25, 2012. 5
- [22] Balaji Lakshminarayanan, Alexander Pritzel, and Charles Blundell. Simple and scalable predictive uncertainty estimation using deep ensembles. *Advances in neural information processing systems*, 30, 2017. 3
- [23] Yann Le and Xuan Yang. Tiny imagenet visual recognition challenge. *CS 231N*, 7(7):3, 2015. 5
- [24] Kimin Lee, Kibok Lee, Honglak Lee, and Jinwoo Shin. A simple unified framework for detecting out-of-distribution samples and adversarial attacks. *Advances in neural information processing systems*, 31, 2018. 2, 5
- [25] Shiyu Liang, Yixuan Li, and Rayadurgam Srikant. Enhancing the reliability of out-of-distribution image detection in neural networks. *arXiv preprint arXiv:1706.02690*, 2017. 2, 3, 5

- [26] Chaoyue Liu and Mikhail Belkin. Clustering with bregman divergences: an asymptotic analysis. *Advances in neural information processing systems*, 29, 2016. 1
- [27] 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. 2, 3, 5
- [28] Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu. Towards deep learning models resistant to adversarial attacks. *arXiv preprint arXiv:1706.06083*, 2017. 2
- [29] Hossein Mirzaei, Ali Ansari, Bahar Dibaei Nia, Mojtaba Nafez, Moein Madadi, Sepehr Rezaee, Zeinab Taghavi, Arad Maleki, Kian Shamsaie, Mahdi Hajjalilue, et al. Scanning trojaned models using out-of-distribution samples. *Advances in Neural Information Processing Systems*, 37:132545–132582, 2025. 1
- [30] Pavlo Molchanov, Arun Mallya, Stephen Tyree, Iuri Frosio, and Jan Kautz. Importance estimation for neural network pruning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 11264–11272, 2019. 3
- [31] Peyman Morteza and Yixuan Li. Provable guarantees for understanding out-of-distribution detection. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 7831–7840, 2022. 2
- [32] Fangzhou Mu, Yingyu Liang, and Yin Li. Gradients as features for deep representation learning. *arXiv preprint arXiv:2004.05529*, 2020. 2
- [33] Yuval Netzer, Tao Wang, Adam Coates, Alessandro Bisacco, Baolin Wu, Andrew Y Ng, et al. Reading digits in natural images with unsupervised feature learning. In *NIPS workshop on deep learning and unsupervised feature learning*, page 4. Granada, 2011. 5
- [34] Bo Peng, Yadan Luo, Yonggang Zhang, Yixuan Li, and Zhen Fang. Conjnorm: Tractable density estimation for out-of-distribution detection. *arXiv preprint arXiv:2402.17888*, 2024. 2, 5
- [35] Vikash Sehwal, Mung Chiang, and Prateek Mittal. Ssd: A unified framework for self-supervised outlier detection. *arXiv preprint arXiv:2103.12051*, 2021. 2
- [36] Yiyu Sun and Yixuan Li. Dice: Leveraging sparsification for out-of-distribution detection. In *European conference on computer vision*, pages 691–708. Springer, 2022. 2
- [37] Yiyu Sun, Chuan Guo, and Yixuan Li. React: Out-of-distribution detection with rectified activations. *Advances in neural information processing systems*, 34:144–157, 2021. 2, 3, 5
- [38] Yiyu Sun, Yifei Ming, Xiaojin Zhu, and Yixuan Li. Out-of-distribution detection with deep nearest neighbors. In *International Conference on Machine Learning*, pages 20827–20840. PMLR, 2022. 2, 3, 5
- [39] Hao Wu, Changhu Wang, Fan Xu, Jinbao Xue, Chong Chen, Xian-Sheng Hua, and Xiao Luo. Pure: Prompt evolution with graph ode for out-of-distribution fluid dynamics modeling. *Advances in Neural Information Processing Systems*, 37:104965–104994, 2025. 1
- [40] Jianxiong Xiao, James Hays, Krista A Ehinger, Aude Oliva, and Antonio Torralba. Sun database: Large-scale scene recognition from abbey to zoo. In *2010 IEEE computer society conference on computer vision and pattern recognition*, pages 3485–3492. IEEE, 2010. 5
- [41] Pingmei Xu, Krista A Ehinger, Yinda Zhang, Adam Finkelstein, Sanjeev R Kulkarni, and Jianxiong Xiao. Turkergaze: Crowdsourcing saliency with webcam based eye tracking. *arXiv preprint arXiv:1504.06755*, 2015. 5
- [42] Jinggang Yang, Kaiyang Zhou, Yixuan Li, and Ziwei Liu. Generalized out-of-distribution detection: A survey. *International Journal of Computer Vision*, 132(12):5635–5662, 2024. 1
- [43] Fisher Yu, Ari Seff, Yinda Zhang, Shuran Song, Thomas Funkhouser, and Jianxiong Xiao. Lsun: Construction of a large-scale image dataset using deep learning with humans in the loop. *arXiv preprint arXiv:1506.03365*, 2015. 5
- [44] Bolei Zhou, Agata Lapedriza, Aditya Khosla, Aude Oliva, and Antonio Torralba. Places: A 10 million image database for scene recognition. *IEEE transactions on pattern analysis and machine intelligence*, 40(6):1452–1464, 2017. 5