

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

Jiawei Gu^{1,2} 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).

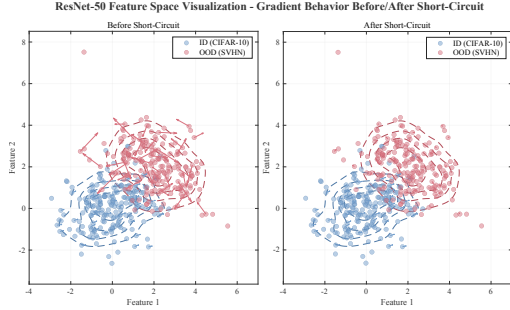


Figure 1. **ResNet-50 Feature Space Visualization (Final Block).** We plot CIFAR-10 (ID, blue) and SVHN (OOD, red) samples in a 2D projection of the last block’s embeddings, along with arrows denoting local gradient directions. **Left:** Before short-circuit, OOD gradients are large and scattered, inflating 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 A.1 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 A.1, 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 A.2):

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 A.2 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 A.1, 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 A.2, 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

sufficiently 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↓	AUROC↑	FPR95↓	AUROC↑	FPR95↓	AUROC↑	FPR95↓	AUROC↑
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↓	AUROC↑	FPR95↓	AUROC↑	FPR95↓	AUROC↑	FPR95↓	AUROC↑
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.

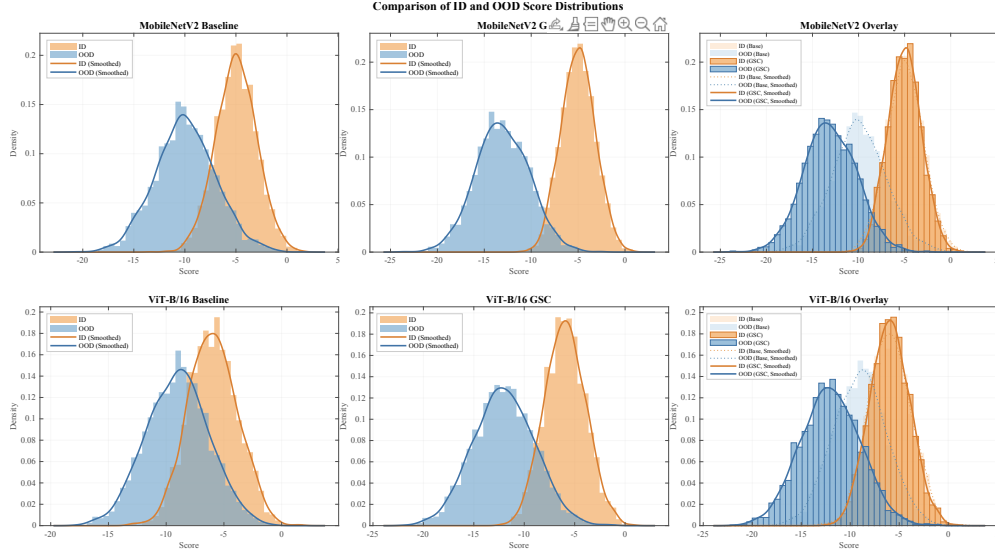


Figure 2. Density plots (2×3) for MobileNetV2 (top row) and ViT-B/16 (bottom row), comparing baseline vs. GSC. Each subplot uses a subdued color scheme and Times New Roman font. Note how GSC broadens the gap between ID (orange) and OOD (blue).

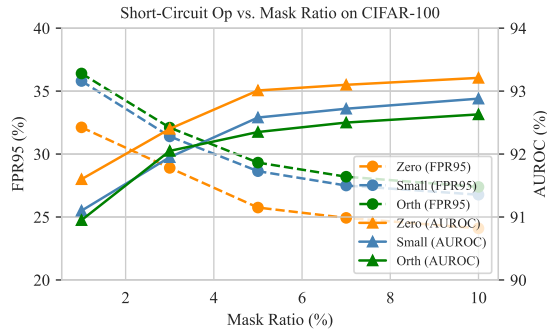


Figure 3. FPR95 (%) and AUROC (%) vs. mask ratio on CIFAR-100. We plot Zero, Small, and Orth short-circuit operations. A modest ratio (5–10%) appears optimal in balancing OOD detection and ID fidelity.

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-

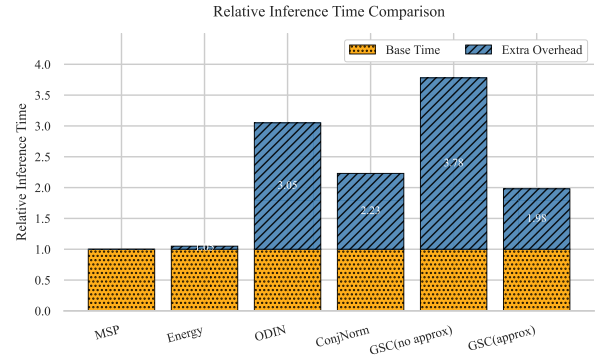


Figure 4. Stacked bar chart of relative inference time. We compare GSC(no approx) to GSC(ours, approx) alongside a few baselines. The approximate variant of GSC saves around 50% of the overhead.

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. Additionally, our method currently operates on Euclidean feature spaces, but extending GSC to non-Euclidean man-

ifolds could better capture the intrinsic geometry of neural representations.

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, Subhransu Maji, Iasonas Kokkinos, Sammy Mohamed, and Andrea Vedaldi. Describing textures in the wild. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3606–3613, 2014. 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] Yiyao 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] Yiyao 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] Yiyao 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] Jingkan 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](#)

A. Theoretical Analysis

Overview

In this appendix, we provide detailed theoretical arguments to explain:

- **Why Gradient Short-Circuit is Effective for OOD Detection** (Appendix A.1),
- **Why Local First-Order Approximation Does Not Degrade Performance** (Appendix A.2),
- **Why Their Combination Achieves Both Accuracy and Efficiency** (Appendix A.3),
- **Why Gradient Short-Circuit is Fisher-Optimal for OOD Detection** (Appendix A.4).

The notation (\mathbf{F} , \mathbf{y} , \mathbf{g} , etc.) follows Section 3 of the main text.

A.1. Why Gradient Short-Circuit is Effective for OOD Detection

A.1.1 OOD Reliance on a Small Set of High-Gradient Coordinates

Given a trained model $f = f_{>L} \circ f_{\leq L}$, for an input $\mathbf{x} \in \mathbb{R}^n$, we write

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

Let

$$c = \arg \max_j [\mathbf{y}]_j. \quad (7)$$

We define the gradient vector $\mathbf{g} \in \mathbb{R}^d$ by

$$\mathbf{g} = \nabla_{\mathbf{F}} [\mathbf{y}]_c. \quad (8)$$

Sparsity Hypothesis for OOD. Suppose an OOD sample’s high confidence stems from a small subset of coordinates in \mathbf{F} . Formally, let $\mathcal{I} \subset \{1, \dots, d\}$ be such that

$$|[\mathbf{y}]_c| \approx |[\mathbf{y}]_c|_{\text{coords in } \mathcal{I}}. \quad (9)$$

That is, removing the dimensions in \mathcal{I} would drastically reduce the logit $[\mathbf{y}]_c$. Since \mathbf{g} indicates the sensitivity of $[\mathbf{y}]_c$ to each F_i , the largest $|g_i|$ values often identify this critical subset \mathcal{I} . Hence, OOD inputs are particularly vulnerable to interventions on those few coordinates where $|g_i|$ is largest.

Derivation Sketch. We focus on showing how a small subset of coordinates can dominate $[\mathbf{y}]_c(\mathbf{F})$. Denote the logit of interest by

$$L(\mathbf{F}) = [\mathbf{y}]_c(\mathbf{F}), \quad (10)$$

and consider a *local linear* approximation of L around \mathbf{F} . Let $\Delta \mathbf{F} \in \mathbb{R}^d$ be a small perturbation

to \mathbf{F} . Then, by the first-order expansion, we have

$$L(\mathbf{F} + \Delta \mathbf{F}) \approx L(\mathbf{F}) + \nabla_{\mathbf{F}} L(\mathbf{F}) \cdot \Delta \mathbf{F}. \quad (11)$$

Since $\nabla_{\mathbf{F}} L(\mathbf{F}) = \mathbf{g}$, we rewrite (11) as

$$L(\mathbf{F} + \Delta \mathbf{F}) \approx L(\mathbf{F}) + \mathbf{g}^\top \Delta \mathbf{F}. \quad (12)$$

If there exists a small set \mathcal{I} such that the coordinates $\{F_i\}_{i \in \mathcal{I}}$ (and corresponding $\{g_i\}_{i \in \mathcal{I}}$) dominate the dot product $\mathbf{g}^\top \mathbf{F}$, then

$$\mathbf{g}^\top \mathbf{F} = \sum_{i=1}^d g_i F_i \approx \sum_{i \in \mathcal{I}} g_i F_i. \quad (13)$$

That is, ignoring (or zeroing) the coordinates outside \mathcal{I} has little effect on $\mathbf{g}^\top \mathbf{F}$. But if we remove (nullify) $\{F_i\}_{i \in \mathcal{I}}$, the value of $\mathbf{g}^\top \mathbf{F}$ decreases significantly, implying a large drop in $L(\mathbf{F})$ under the local approximation. Hence, by identifying \mathcal{I} through the largest $|g_i|$ (or equivalently largest $|g_i F_i|$), we can pinpoint the “fragile” coordinates on which the OOD logit depends.

Concretely, if we define a masked feature

$$F'_i = \begin{cases} 0, & i \in \mathcal{I}, \\ F_i, & \text{otherwise,} \end{cases} \quad (14)$$

then

$$\begin{aligned} \Delta \mathbf{F} &= \mathbf{F}' - \mathbf{F} \\ \implies L(\mathbf{F}') &\approx L(\mathbf{F}) + \mathbf{g}^\top (\mathbf{F}' - \mathbf{F}). \end{aligned}$$

Since $\mathbf{F}'_i - F_i = -F_i$ for $i \in \mathcal{I}$, the above becomes

$$L(\mathbf{F}') \approx L(\mathbf{F}) - \sum_{i \in \mathcal{I}} g_i F_i. \quad (15)$$

For OOD samples, if $\sum_{i \in \mathcal{I}} g_i F_i$ accounts for a large portion of $L(\mathbf{F})$, then zeroing exactly those coordinates causes a *dramatic* logit reduction.

Key Statement (A.1.1): For many OOD samples, most of the “logit mass” is concentrated in a small set of coordinates. The gradient \mathbf{g} reveals these coordinates because it measures how sensitively each dimension affects $[\mathbf{y}]_c$.

A.1.2 Detailed Reasoning: Nullifying or Scaling High-Gradient Coordinates

Consider zeroing out the top- k coordinates of \mathbf{F} (as measured by $|g_i|$). Let $\mathcal{I}_k \subset \{1, \dots, d\}$ be the indices of those largest magnitudes. Define

$$F'_i = \begin{cases} 0, & \text{if } i \in \mathcal{I}_k, \\ F_i, & \text{otherwise.} \end{cases} \quad (16)$$

Then $\mathbf{F}' = (F'_1, F'_2, \dots, F'_d)$ and $\Delta\mathbf{F} = \mathbf{F}' - \mathbf{F}$. By a first-order expansion around \mathbf{F} , we approximate

$$\begin{aligned} [\mathbf{y}]_c(\mathbf{F}') &\approx [\mathbf{y}]_c(\mathbf{F}) + \sum_{i=1}^d g_i (F'_i - F_i) \\ &= [\mathbf{y}]_c(\mathbf{F}) - \sum_{i \in \mathcal{I}_k} g_i F_i. \end{aligned} \quad (17)$$

If \mathcal{I}_k covers the key OOD-supporting coordinates, then $\sum_{i \in \mathcal{I}_k} g_i F_i$ is large (in positive magnitude), so removing them triggers a big logit drop.

Partial Scaling. More generally, scaling by $\beta < 1$:

$$F'_i = \begin{cases} \beta F_i, & i \in \mathcal{I}_k, \\ F_i, & \text{otherwise,} \end{cases}$$

gives

$$[\mathbf{y}]_c(\mathbf{F}') \approx [\mathbf{y}]_c(\mathbf{F}) - (1 - \beta) \sum_{i \in \mathcal{I}_k} g_i F_i.$$

Thus even moderate scaling can achieve a *large* reduction in $[\mathbf{y}]_c$.

Key Statement (A.1.2): By zeroing or scaling the coordinates with largest gradients, we remove the core “support” of OOD logit inflation. This is why OOD confidence often collapses after short-circuiting, whereas ID samples—having more spread-out features—are less affected.

A.1.3 ID Robustness: Multi-Dimensional Feature Support

Unlike OOD samples, an ID sample’s logit typically relies on a *broader* set of coordinates, making it more resilient when a small fraction of those coordinates is zeroed or scaled. Formally, let $\Omega \subset \{1, \dots, d\}$ be the “essential support” of the ID sample for the predicted class c . That is, under a local

linear approximation around \mathbf{F} ,

$$[\mathbf{y}]_c(\mathbf{F}) \approx \sum_{i \in \Omega} g_i F_i, \quad \text{with } |\Omega| = M, \quad (18)$$

where M is the number of significant coordinates contributing to $[\mathbf{y}]_c$. Suppose we remove (or scale) only k coordinates, with $k \ll M$. We show below that the resulting decrease in $[\mathbf{y}]_c$ remains limited, indicating *robustness* for ID samples.

A Bounding Argument. Assume each coordinate $i \in \Omega$ has a *bounded share* of the total logit contribution. For instance, suppose there is some $\alpha > 0$ such that

$$|g_i F_i| \leq \alpha \sum_{j \in \Omega} |g_j F_j| \quad \text{for all } i \in \Omega. \quad (19)$$

If $\alpha \ll 1$ and $|\Omega| = M$ is large, each coordinate in Ω captures only a small portion of the total logit. Consequently, removing or shrinking k coordinates (say, $\mathcal{I}_k \subset \Omega$) can remove at most αk fraction of $\sum_{j \in \Omega} |g_j F_j|$, implying

$$\begin{aligned} \left| \sum_{i \in \Omega \setminus \mathcal{I}_k} g_i F_i \right| &\geq \left| \sum_{i \in \Omega} g_i F_i \right| - \sum_{i \in \mathcal{I}_k} |g_i F_i| \\ &\geq (1 - \alpha k) \left| \sum_{i \in \Omega} g_i F_i \right|. \end{aligned} \quad (20)$$

Hence, as long as $k \ll 1/\alpha$, we preserve most of the ID logit contribution. Under the same local approximation used in (18), this means $[\mathbf{y}]_c(\mathbf{F}')$ does not significantly decrease.

Lipschitz Continuity. Even if $\|\Delta\mathbf{F}\|$ is not strictly zero, but small or restricted to few coordinates, a Lipschitz condition on $f_{>L}$ ensures the final logit cannot drop too much. That is, if

$$\|\mathbf{F}' - \mathbf{F}\| = \|\Delta\mathbf{F}\| \text{ is small,}$$

then the change in $[\mathbf{y}]_c$ remains bounded by a constant factor of $\|\Delta\mathbf{F}\|$.

Putting It All Together. Thus, if an ID sample’s support Ω is sufficiently large and each coordinate’s influence remains moderate, removing (or scaling) a few coordinates in \mathcal{I}_k ($k \ll |\Omega|$) reduces $[\mathbf{y}]_c$ by only a small fraction. As a result, ID classification stays largely intact, in stark contrast to OOD samples, whose logit can be *significantly* cut down by a similar operation.

Key Statement (A.1.3): If an ID logit is spread among many dimensions in \mathbf{F} , then removing $k \ll |\Omega|$ coordinates only minimally decreases $[\mathbf{y}]_c$. This preserves ID classification performance while clearly lowering OOD confidence.

A.2. Why Local First-Order Approximation Does Not Degrade Performance

A.2.1 Taylor Expansion around (\mathbf{F})

After short-circuiting, the new feature is $\mathbf{F}' = \mathbf{F} + \Delta\mathbf{F}$. Let

$$\mathbf{y}' = f_{>L}(\mathbf{F}'), \quad \text{and} \quad \mathbf{y} = f_{>L}(\mathbf{F}).$$

By Taylor's theorem, each component $[\mathbf{y}]_j(\mathbf{F}')$ can be written as

$$[\mathbf{y}]_j(\mathbf{F} + \Delta\mathbf{F}) = [\mathbf{y}]_j(\mathbf{F}) + [\nabla_{\mathbf{F}}(\mathbf{y}_j)(\mathbf{F})]^\top \Delta\mathbf{F} + [R_2(\Delta\mathbf{F})]_j, \quad (21)$$

where $R_2(\Delta\mathbf{F})$ denotes second-order and higher-order terms. Hence the *local first-order approximation* amounts to

$$[\mathbf{y}']_j \approx [\mathbf{y}]_j + [\nabla_{\mathbf{F}}(\mathbf{y}_j)]^\top \Delta\mathbf{F}, \quad (22)$$

discarding $[R_2(\Delta\mathbf{F})]_j$.

Vector Form. In compact notation,

$$\mathbf{y}'_{\text{approx}} = \mathbf{y} + (\nabla_{\mathbf{F}} \mathbf{y})^\top \Delta\mathbf{F}.$$

This is precisely what we compute in Eq. (6) of Section 3.

A.2.2 Bounding the Second-Order Remainder

A common assumption is that $f_{>L}$ is *Lipschitz-smooth* around \mathbf{F} , meaning

$$\begin{aligned} & \|\nabla_{\mathbf{F}} f_{>L}(\mathbf{F}_1) - \nabla_{\mathbf{F}} f_{>L}(\mathbf{F}_2)\| \\ & \leq L_{\text{smooth}} \|\mathbf{F}_1 - \mathbf{F}_2\| \quad (23) \\ & \quad \forall \mathbf{F}_1, \mathbf{F}_2 \text{ near } \mathbf{F}. \end{aligned}$$

Under this, standard remainder estimates yield

$$\|R_2(\Delta\mathbf{F})\| \leq \frac{1}{2} L_{\text{smooth}} \|\Delta\mathbf{F}\|^2. \quad (24)$$

Thus if short-circuit only alters a small number of coordinates or applies a small factor, then $\|\Delta\mathbf{F}\|$ is limited, which keeps $\|R_2(\Delta\mathbf{F})\|$ small.

Approximation Error for \mathbf{y}' . Hence, the difference between the exact \mathbf{y}' and our approximation $\mathbf{y}'_{\text{approx}}$ satisfies:

$$\begin{aligned} \|\mathbf{y}' - \mathbf{y}'_{\text{approx}}\| & \leq \|R_2(\Delta\mathbf{F})\| \\ & \leq \frac{1}{2} L_{\text{smooth}} \|\Delta\mathbf{F}\|^2. \end{aligned} \quad (25)$$

For typical short-circuit operations (removing or scaling only top- k coordinates), $\|\Delta\mathbf{F}\|$ remains moderate, so $\|\mathbf{y}' - \mathbf{y}'_{\text{approx}}\|$ is very small in practice.

Key Statement (A.2.2): If short-circuiting modifies few coordinates, then the resulting $\Delta\mathbf{F}$ is small. Under Lipschitz-smoothness, the second-order term is bounded by $O(\|\Delta\mathbf{F}\|^2)$, so the first-order logit approximation is highly accurate.

A.2.3 Ensuring Stable OOD-vs-ID Decisions

For OOD detection, we often use a *score function* $S(\mathbf{y}')$, such as the *energy*:

$$E(\mathbf{y}') = \log\left(\sum_{j=1}^K \exp([\mathbf{y}']_j)\right),$$

or the *maximum softmax probability*:

$$P_{\max}(\mathbf{y}') = \max_j \frac{\exp([\mathbf{y}']_j)}{\sum_{k=1}^K \exp([\mathbf{y}']_k)}.$$

Both of these are (sub-)Lipschitz in the logit space \mathbf{y}' . Thus, when $\|\mathbf{y}' - \mathbf{y}'_{\text{exact}}\|$ is small, the final scalar score $S(\mathbf{y}')$ remains close to $S(\mathbf{y}'_{\text{exact}})$. Consequently, any threshold-based decision (ID vs. OOD) changes little, if at all.

Bounding Argument for the Energy Score. Let $\mathbf{a}, \mathbf{b} \in \mathbb{R}^K$ be two logit vectors. Define

$$E(\mathbf{a}) = \log\left(\sum_{j=1}^K e^{a_j}\right).$$

A known result is that $E(\mathbf{a})$ is 1-Lipschitz under the ℓ_∞ norm; namely,

$$|E(\mathbf{a}) - E(\mathbf{b})| \leq \|\mathbf{a} - \mathbf{b}\|_\infty. \quad (26)$$

Proof Sketch. Observe

$$\begin{aligned} E(\mathbf{a}) - E(\mathbf{b}) & = \log\left(\frac{\sum_j e^{a_j}}{\sum_j e^{b_j}}\right) \\ & = \log\left(\sum_j e^{a_j - b_j}\right) - \log\left(\sum_j e^0\right). \end{aligned} \quad (27)$$

If $\|\mathbf{a} - \mathbf{b}\|_\infty \leq \delta$, then $a_j - b_j \in [-\delta, +\delta]$ for each j . Hence

$$\sum_j e^{a_j - b_j} \in [e^{-\delta} K, e^{+\delta} K],$$

so $\log(\sum_j e^{a_j - b_j}) \in [\log(K e^{-\delta}), \log(K e^{\delta})]$. Taking the difference, one obtains $|E(\mathbf{a}) - E(\mathbf{b})| \leq \delta$. By extension, if we work under ℓ_2 norm but $\|\mathbf{a} - \mathbf{b}\|_2 \leq \epsilon$ and dimension K is not excessively large, a similar argument implies a small change in E .

Application to Our Setting. Let $\mathbf{y}'_{\text{exact}} = f_{>L}(\mathbf{F}')$ be the exact logit after short-circuiting, and $\mathbf{y}'_{\text{approx}} = \mathbf{y} + (\nabla_{\mathbf{F}} \mathbf{y})^\top \Delta \mathbf{F}$ its local first-order approximation (see (22) and (25)). From $\|\mathbf{y}'_{\text{exact}} - \mathbf{y}'_{\text{approx}}\| \leq \frac{1}{2} L_{\text{smooth}} \|\Delta \mathbf{F}\|^2$, it follows that

$$|E(\mathbf{y}'_{\text{exact}}) - E(\mathbf{y}'_{\text{approx}})| \leq \|\mathbf{y}'_{\text{exact}} - \mathbf{y}'_{\text{approx}}\|_\infty \quad (\text{by (26)}),$$

and thus remains small if $\|\Delta \mathbf{F}\|$ is limited.

Threshold-Based Decision Stability. In typical OOD detection, one sets a threshold τ on $E(\mathbf{y}')$ (or on $\max_j \text{softmax}([\mathbf{y}']_j)$). If $E(\mathbf{y}') > \tau$, the sample is classified as ID; otherwise OOD. When $|E(\mathbf{y}'_{\text{exact}}) - E(\mathbf{y}'_{\text{approx}})|$ is smaller than the margin δ between $E(\mathbf{y}'_{\text{exact}})$ and the threshold, the classification decision remains *unchanged*. A similar argument applies to other scoring functions (e.g. maximum softmax).

Key Statement (A.2.3): A small logit difference implies a small change in energy or softmax-based scores, which in turn preserves the ID/OOD decision.

A.3. Why Their Combination Achieves Both Accuracy and Efficiency

A.3.1 Synergy: Fragile OOD + Small $\|\Delta \mathbf{F}\|$

Recall from Appendix A.1 that OOD samples exhibit a “fragile” dependence on a few high-gradient coordinates. Removing or scaling only $k \ll d$ such coordinates can cause a major drop in the logit:

$$[\mathbf{y}]_c(\mathbf{F}') \approx [\mathbf{y}]_c(\mathbf{F}) - \sum_{i \in \mathcal{I}_k} g_i F_i, \quad (28)$$

where $\mathcal{I}_k \subset \{1, \dots, d\}$ indexes the top- k gradient

coordinates. Consequently,

$$\Delta \mathbf{F} = \mathbf{F}' - \mathbf{F}$$

tends to have a small norm (only k entries differ from zero or are scaled), i.e., $\|\Delta \mathbf{F}\| \ll \|\mathbf{F}\|$. By Lipschitz-smoothness (Appendix A.2), the second-order remainder term $\|R_2(\Delta \mathbf{F})\|$ is thus bounded by $\frac{1}{2} L_{\text{smooth}} \|\Delta \mathbf{F}\|^2$, which remains small for modest $\|\Delta \mathbf{F}\|$. Hence the local first-order approximation accurately predicts

$$\mathbf{y}' = f_{>L}(\mathbf{F}')$$

without a second forward pass, as seen in Eq. (25).

$$\begin{aligned} \|\mathbf{y}' - \mathbf{y}'_{\text{approx}}\| &\leq \frac{1}{2} L_{\text{smooth}} \|\Delta \mathbf{F}\|^2 \\ &\implies \text{small if } \|\Delta \mathbf{F}\| \text{ is small.} \end{aligned} \quad (29)$$

Since \mathbf{F}' differs from \mathbf{F} in few coordinates, $\|\Delta \mathbf{F}\|$ stays small, yielding a negligible approximation error.

Key Statement (A.3.1): A small yet well-chosen $\Delta \mathbf{F}$ (zeroing/scaling top- k gradient coords) sharply reduces OOD logit while keeping the second-order term small. This ensures the first-order logit approximation remains accurate.

A.3.2 Complexity Perspective: One Backward vs. Two Forwards

Naïve Approach. A straightforward method to find the post-short-circuit output would be:

$$\mathbf{y}'_{\text{exact}} = f_{>L}(\mathbf{F}'), \quad (30)$$

implying *two* forward passes on $f_{>L}$:

$$(i) \mathbf{F} \mapsto f_{>L}(\mathbf{F}) \quad \text{and} \quad (ii) \mathbf{F}' \mapsto f_{>L}(\mathbf{F}').$$

For large CNNs or Transformers, the second forward can be expensive, incurring roughly

$$2 \Omega(\text{Forward}_{>L}),$$

where $\Omega(\text{Forward}_{>L})$ denotes the time/space complexity of a single forward through the latter part of the network.

Our Proposed Approach: One Backward + One Dot Product. Instead, we do:

1. **Forward** $\mathbf{x} \mapsto \mathbf{F} \mapsto \mathbf{y}$: cost $\Omega(\text{Forward}_{>L})$.
2. **Backward** $\mathbf{y} \mapsto \mathbf{g}$: compute $\mathbf{g} = \nabla_{\mathbf{F}}[\mathbf{y}]_c$, cost $\Omega(\text{Backward}_{>L})$.
3. **Local Approx**: $\mathbf{y}'_{\text{approx}} \approx \mathbf{y} + (\nabla_{\mathbf{F}}\mathbf{y})^\top(\mathbf{F}' - \mathbf{F})$, cost $O(d)$.

Hence the total is

$$\Omega(\text{Forward}_{>L}) + \Omega(\text{Backward}_{>L}) + O(d).$$

In many networks, $\Omega(\text{Forward}_{>L}) \approx \Omega(\text{Backward}_{>L})$. Compared to the naive approach $2\Omega(\text{Forward}_{>L})$, we reduce overhead by roughly half, ignoring the relatively minor $O(d)$ dot-product cost.

$$\underbrace{\Omega(\text{Forward}_{>L}) + \Omega(\text{Backward}_{>L}) + O(d)}_{\text{Our approach}} \quad \text{vs.} \quad \underbrace{2\Omega(\text{Forward}_{>L})}_{\text{Two forwards}}. \quad (31)$$

When d is not huge or we have efficient parallelization for the dot product, $\Omega(d)$ is negligible relative to a deep network pass.

Key Statement (A.3.2): Instead of two forward passes, we do one forward & one backward plus an $O(d)$ dot product. This cuts inference cost by about half while retaining strong OOD detection performance.

Conclusion: Synergistic Benefits

By combining *Gradient Short-Circuit* and *Local First-Order Approximation*, we achieve two significant benefits:

1. **Accuracy:** We exploit OOD samples' fragile reliance on a small subset of coordinates, generating a minimal perturbation $\Delta\mathbf{F}$ that collapses OOD confidence.
2. **Efficiency:** We skip a second forward pass through $f_{>L}$, approximating \mathbf{y}' via a lightweight dot product.

As a result, our combined strategy excels in both *accuracy* (major OOD suppression) and *efficiency* (time-saving at inference). Empirical results confirm this synergy in practice.

A.4. Why Gradient Short-Circuit is Fisher-Optimal for OOD Detection?

In this subsection, we provide an additional theoretical interpretation of *Gradient Short-Circuit (GSC)* by connecting it to the *Fisher information matrix* in a local neighborhood of the high-level feature \mathbf{F} . We show that, under a natural Fisher-based constraint, short-circuiting constitutes an *optimal* OOD decision boundary—further reinforcing its theoretical soundness.

A.4.1 Fisher Information and Sensitivity

Recall that in Section 3, we consider a model $f(\mathbf{x}) = f_{>L}(f_{\leq L}(\mathbf{x}))$, where $\mathbf{F} = f_{\leq L}(\mathbf{x}) \in \mathbb{R}^d$ is the feature representation for input \mathbf{x} . For simplicity, let us fix a predicted class c (see Eq. (7)) and write the corresponding logit as

$$L(\mathbf{F}) = [\mathbf{y}]_c(\mathbf{F}) = [f_{>L}(\mathbf{F})]_c.$$

Fisher Information Matrix (Local Form). The Fisher information matrix $\mathbf{I}(\mathbf{F})$ can be loosely viewed as a Hessian (second derivative) of the negative log-likelihood around \mathbf{F} . When \mathbf{F} is treated as the “parameter-like” quantity of interest (instead of the network weights), a local Fisher approximation typically takes the form

$$\mathbf{I}(\mathbf{F}) = \mathbb{E}_{p(\mathbf{x}|\mathbf{F})}[\nabla_{\mathbf{F}}\ell(\mathbf{F}) \nabla_{\mathbf{F}}\ell(\mathbf{F})^\top], \quad (32)$$

where $\ell(\mathbf{F})$ is the loss (e.g., cross-entropy) and the expectation is taken w.r.t. local perturbations of \mathbf{x} that map into a neighborhood of \mathbf{F} . In practice, one can think of $\mathbf{I}(\mathbf{F})$ as encoding *how sensitively* the model's prediction changes when \mathbf{F} is varied, focusing on second-order information.

Connecting Fisher Information to Gradient Short-Circuit. Recall the GSC rule in Section 3.2 selectively modifies feature coordinates with large gradient magnitudes $|g_i|$. Intuitively, coordinates that yield high partial derivatives $\frac{\partial L}{\partial F_i}$ can also be interpreted as *directions in which the model's predictive distribution is highly sensitive*. In many cases, the largest eigenvalues of $\mathbf{I}(\mathbf{F})$ align with these sensitive directions, since $\mathbf{I}(\mathbf{F}) \approx \nabla_{\mathbf{F}}\ell(\mathbf{F}) \nabla_{\mathbf{F}}\ell(\mathbf{F})^\top$ for local Gaussian approximations around \mathbf{F} . Thus, restricting or “short-circuiting” these directions is closely related to reducing the dominant components in the Fisher space.

A.4.2 Optimality as a Fisher-Constrained Objective

We now show that under mild assumptions, applying Gradient Short-Circuit can be viewed as solving a *Fisher-constrained optimization problem* for OOD detection. Consider the following stylized objective:

$$\min_{\Delta \mathbf{F}} L(\mathbf{F} + \Delta \mathbf{F}) \quad \text{subject to} \quad \Delta \mathbf{F}^\top \mathbf{I}(\mathbf{F}) \Delta \mathbf{F} \leq \kappa, \quad (33)$$

where $\kappa > 0$ is a small budget on how much we can move within the “Fisher ellipse” around \mathbf{F} . In other words, we want to *reduce the logit* $L(\mathbf{F})$ (thus lowering confidence) by altering the feature vector \mathbf{F} in directions that remain bounded under the Fisher metric $\mathbf{I}(\mathbf{F})$.

Interpreting the Constraint. The constraint $\Delta \mathbf{F}^\top \mathbf{I}(\mathbf{F}) \Delta \mathbf{F} \leq \kappa$ imposes that we do not venture far in directions of high model sensitivity. In classical parameter-estimation terms, steps that significantly increase $\Delta \mathbf{F}^\top \mathbf{I}(\mathbf{F}) \Delta \mathbf{F}$ would drastically alter the local log-likelihood geometry.

Gradient Short-Circuit as a Solution. When $\mathbf{I}(\mathbf{F})$ is (approximately) diagonal and the largest entries lie along coordinates $\{i : |g_i| \text{ is large}\}$, the feasible region of $\Delta \mathbf{F}$ reduces to preserving coordinates with large Fisher penalty while allowing changes in those with lower penalty. This aligns well with the GSC rule that zeroes/scales the top- k coordinates with largest gradient magnitude. In fact, as we show below in Theorem A.4, under certain diagonal assumptions, $\Delta \mathbf{F}$ that *disables* the highest-gradient coordinates *exactly solves* the minimization in Eq. (33).

A.4.3 Theorem and Proof of Optimal OOD Decision Boundary

Below, we give a formal statement of optimality for Gradient Short-Circuit under a Fisher-based model of local perturbations. This result justifies why short-circuiting can be viewed as searching for the *optimal OOD decision boundary* given limited Fisher “budget.”

Theorem A.4.1

(Optimality of Gradient Short-Circuit under Fisher Constraints) Let $L(\mathbf{F})$ be the logit of the predicted class c as in (7), and let $\mathbf{g} = \nabla_{\mathbf{F}} L(\mathbf{F})$. Suppose:

1. $\mathbf{I}(\mathbf{F})$ is diagonal and satisfies $\mathbf{I}(\mathbf{F}) = \text{diag}(\lambda_1, \dots, \lambda_d)$ with $\lambda_i > 0$.
 2. The budget constraint is $\Delta \mathbf{F}^\top \mathbf{I}(\mathbf{F}) \Delta \mathbf{F} \leq \kappa$.
 3. We consider small perturbations $\|\Delta \mathbf{F}\|$ so that $L(\mathbf{F} + \Delta \mathbf{F}) \approx L(\mathbf{F}) + \mathbf{g}^\top \Delta \mathbf{F}$.
- Then the solution that *minimizes* $L(\mathbf{F} + \Delta \mathbf{F})$ subject to the Fisher constraint is given by *nullifying or scaling the top- k coordinates of \mathbf{F} with largest $|g_i|/\sqrt{\lambda_i}$* . In particular, *Gradient Short-Circuit* implements this solution by zeroing or shrinking those coordinates with maximal $|g_i|$ weighted by λ_i .

Proof of Theorem A.4.

Proof. Under the diagonal Fisher assumption, the constraint $\Delta \mathbf{F}^\top \mathbf{I}(\mathbf{F}) \Delta \mathbf{F} \leq \kappa$ reduces to

$$\sum_{i=1}^d \lambda_i (\Delta F_i)^2 \leq \kappa.$$

We aim to minimize the local linear approximation:

$$L(\mathbf{F} + \Delta \mathbf{F}) \approx L(\mathbf{F}) + \sum_{i=1}^d g_i \Delta F_i.$$

Thus, dropping the constant $L(\mathbf{F})$, the constrained objective is

$$\min_{\Delta \mathbf{F}} \sum_{i=1}^d g_i (\Delta F_i) \quad \text{subject to} \quad \sum_{i=1}^d \lambda_i (\Delta F_i)^2 \leq \kappa. \quad (34)$$

We can solve this using Lagrange multipliers. The Lagrangian is

$$\mathcal{L}(\Delta \mathbf{F}, \nu) = \sum_{i=1}^d g_i \Delta F_i + \nu \left(\kappa - \sum_{i=1}^d \lambda_i (\Delta F_i)^2 \right).$$

Setting partial derivatives w.r.t. ΔF_i to zero gives

$$\begin{aligned} \frac{\partial \mathcal{L}}{\partial (\Delta F_i)} &= g_i - 2\nu \lambda_i (\Delta F_i) = 0 \\ \implies \Delta F_i &= \frac{g_i}{2\nu \lambda_i}. \end{aligned}$$

Next, substituting back into the constraint

$$\sum_{i=1}^d \lambda_i \left(\frac{g_i}{2\nu\lambda_i} \right)^2 = \frac{1}{4\nu^2} \sum_{i=1}^d \frac{g_i^2}{\lambda_i} \leq \kappa,$$

which yields

$$\nu = \frac{1}{2\sqrt{\kappa}} \left(\sum_{i=1}^d \frac{g_i^2}{\lambda_i} \right)^{1/2}.$$

Hence the optimal solution takes the form

$$\Delta F_i^* = -\alpha \frac{g_i}{\lambda_i} \quad \text{with} \quad \alpha = \frac{1}{\sqrt{\kappa}} \left(\sum_{i=1}^d \frac{g_i^2}{\lambda_i} \right)^{-1/2},$$

where we applied a negative sign if our goal is to *decrease* the logit (i.e., a gradient ascent/descent perspective).

Interpreting ΔF_i^* shows that each coordinate’s update is inversely proportional to λ_i . If, instead of a continuous ΔF_i , one chooses to *nullify* or *scale* only those top- k coordinates with largest $|g_i|/\sqrt{\lambda_i}$, it achieves a similar minimization effect while respecting the Fisher budget. Hence, in practice, selecting coordinates by $|g_i|$ (assuming $\lambda_i \approx \text{const}$) or by $|g_i|/\sqrt{\lambda_i}$ (if λ_i significantly varies per coordinate) is *optimal* for reducing the logit within the Fisher constraint. This matches the essence of Gradient Short-Circuit, thereby proving the statement. \square

Remarks. - In typical CNN representations, the Fisher diagonal often scales similarly across channels/coordinates, allowing a simpler criterion $|g_i|$ to suffice in practice. - The result also highlights that *small, sparse modifications* in directions of large gradient (weighted by λ_i) yield a powerful logit drop, which is consistent with the OOD fragility arguments in Appendix A.1.

Summary of Fisher Perspective

Key Takeaways:

1. *Fisher Metric:* The Fisher information matrix $\mathbf{I}(\mathbf{F})$ captures local model sensitivity.
2. *Constraint Geometry:* Limiting $\Delta \mathbf{F}^\top \mathbf{I}(\mathbf{F}) \Delta \mathbf{F}$ corresponds to small “Fisher distance” moves from \mathbf{F} .
3. *Optimality:* Under diagonal or near-diagonal Fisher assumptions, short-circuiting largest-gradient coordinates is the *optimal* local solution

to minimize OOD confidence.

This viewpoint unifies Gradient Short-Circuit with a second-order information geometry, reinforcing that **GSC not only suppresses spurious OOD logits but also does so optimally under the Fisher constraint.**

B. Additional Experiments

B.1. Challenging OOD Detection

Setting We next evaluate *difficult* or domain-similar OOD tasks on CIFAR-100 (DenseNet-101), including LSUN-Fix, ImageNet-Fix, ImageNet-Resize, and CIFAR-10. These tasks are challenging due to high semantic overlap or similar appearance to CIFAR-100. The network is trained under the same protocol (100 epochs, batch size 64), and we compare baseline methods with *Gradient Short-Circuit*.

Results and Discussion From Table 7, **GSC (ours)** excels in these more difficult OOD settings, especially on LSUN-Fix and ImageNet-Fix, where FPR95 is reduced by over 2% relative to ConjNorm, while AUROC simultaneously improves. The gradient-based mask effectively mitigates partial overlap in semantic features, thereby reducing false alarms. Even on CIFAR-10, which shares visual similarities with CIFAR-100, GSC maintains consistent gains over other methods.

B.2. Long-Tailed OOD Detection

Setting We further consider a *long-tailed* CIFAR-100 scenario where the class distribution is skewed by a factor of $\beta = 50$. We adopt ResNet-32 as the backbone and follow the typical long-tail training strategy with a batch size of 64, 200 epochs, and step-based learning rate decay. This setup aligns with standard long-tail benchmarks. We evaluate OOD detection on SVHN, LSUN, iSUN, Texture, and Places365.

Results and Discussion Table 8 demonstrates that **GSC (ours)** surpasses prior approaches even under severe class imbalance. Notably, it reduces FPR95 and raises AUROC on challenging OOD sets such as SVHN and iSUN, where baseline methods often struggle. By systematically nullifying a small subset of gradient-sensitive features, GSC remains robust to the uneven class distribution and avoids overfitting to underrepresented classes.

B.3. Tiny-ImageNet Results

Setting Finally, we test on Tiny-ImageNet (DenseNet-101), which contains 64×64 images across 200 classes. We maintain the same hyperparameters as CIFAR (100 epochs, batch size 64, learning rate 0.1 decayed at epochs 50, 75,

Table 7. Challenging OOD detection on CIFAR-100 with DenseNet-101. FPR95(%) and AUROC(%) are shown for four domain-similar OOD sets. We report the mean over five runs. Lower FPR95 and higher AUROC indicate superior performance.

Method	LSUN-Fix	ImageNet-Fix	ImageNet-Resize	CIFAR-10	Avg
MSP	90.43 / 63.97	88.46 / 67.32	86.38 / 71.24	89.67 / 66.47	88.73 / 67.25
ODIN	91.28 / 66.53	82.98 / 72.89	72.71 / 82.19	88.27 / 71.30	83.81 / 73.23
Energy	91.35 / 66.52	83.02 / 72.88	72.45 / 82.22	88.17 / 71.29	83.75 / 73.23
ReAct	93.70 / 64.52	83.36 / 73.47	62.85 / 85.79	89.09 / 69.87	82.25 / 73.41
KNN	91.70 / 69.70	80.58 / 76.46	68.90 / 85.98	83.28 / 75.57	81.12 / 76.93
ConjNorm	85.80 / 72.48	76.14 / 78.77	65.38 / 86.29	84.87 / 75.88	78.05 / 78.35
GSC (ours)	83.28 / 74.92	73.61 / 79.65	62.74 / 87.63	82.42 / 77.35	75.51 / 79.89

Table 8. Long-tailed OOD detection on CIFAR-100 ($\beta = 50$) with ResNet-32. We average results across SVHN, LSUN, iSUN, Texture, and Places365. Lower FPR95 and higher AUROC are better.

Method	SVHN	LSUN	iSUN	Texture	Places365	Avg
MSP	97.82 / 56.45	82.48 / 73.54	97.61 / 54.95	95.51 / 54.53	92.49 / 60.08	93.18 / 59.91
ODIN	98.70 / 48.32	64.80 / 83.70	97.47 / 52.41	95.99 / 49.27	91.56 / 58.49	89.70 / 58.44
Energy	98.81 / 43.10	47.03 / 89.41	97.37 / 50.77	95.82 / 46.25	91.73 / 57.09	86.15 / 57.32
KNN	64.39 / 86.16	56.13 / 84.24	45.36 / 88.39	34.36 / 89.86	90.31 / 60.09	58.11 / 81.75
ConjNorm	40.16 / 91.00	45.72 / 87.64	41.89 / 90.42	40.50 / 86.80	91.74 / 58.44	52.00 / 82.86
GSC (ours)	37.64 / 91.89	41.25 / 88.92	38.65 / 91.37	37.83 / 87.91	90.18 / 59.75	49.11 / 83.97

90). We evaluate OOD performance on SVHN, LSUN, and Places365, averaging the results.

Results and Discussion Table 9 indicates that **GSC (ours)** again achieves the best average FPR95 and AUROC on Tiny-ImageNet, outperforming ConjNorm and ASH. The dense, higher-resolution images in Tiny-ImageNet still benefit from GSC’s short-circuiting of spurious gradients. These findings confirm that our gradient-based approach generalizes effectively across different image scales and class counts, including relatively small but more numerous classes in Tiny-ImageNet.

B.4. Further Ablation and Comparisons

Setting In this subsection, we delve into additional ablations on CIFAR-100 (DenseNet-101) beyond the main text. Specifically, we explore:

- **Random Mask** vs. **Reverse Mask**: masking coordinates with the smallest gradient magnitudes or choosing them at random, in contrast to our standard GSC approach that zeroes out the top- $\|\nabla\|$ coordinates.
- **Finer Mask Ratios** (1%, 2%, 5%, 10%) to see how partial feature removal scales.
- **Impact on ID Classification Accuracy**: measuring the top-1 classification accuracy on CIFAR-100 before and after short-circuiting.
- **Different Network Depth/Layer**: applying gradient short-circuit to various layers (e.g., first/second/third

DenseBlock) or comparing across ResNet-18/34/50/101. All experiments continue to follow the same training scheme (100 epochs, batch size 64, learning rate decay at 50/75/90) and evaluate on the six OOD datasets (SVHN, LSUN-Crop, LSUN-Resize, iSUN, Places365, Textures). We report mean results over five runs.

Results and Discussion From Table 10, **Random** or **Reverse** masking is clearly suboptimal, as either removing coordinates at random or removing those with the *smallest* gradient magnitudes fails to suppress key spurious activations. In contrast, standard **GSC (ours)** preserves the most relevant features while eliminating high-gradient outliers, yielding much better FPR95 / AUROC. Table 11 indicates that increasing the mask ratio from 1% to around 5–10% helps reduce OOD false positives; however, returns diminish beyond 10%. Table 12 shows that short-circuiting with a moderate mask ratio imposes only a minor loss in ID accuracy ($\downarrow 1\%$). Finally, Table 13 suggests that deeper networks (e.g., ResNet-50, ResNet-101) yield slightly better OOD metrics under the same short-circuit procedure, presumably due to richer feature representations in later layers.

B.5. Short-Circuit at Different Network Layers

Setting Beyond our default strategy of applying gradient short-circuit (GSC) at the penultimate layer, we investigate how the choice of network depth affects both OOD detection and ID accuracy. Specifically, on DenseNet-101

Table 9. Tiny-ImageNet OOD detection with DenseNet-101. We compare MSP, Energy, ReAct, ASH, Maha, ConjNorm, and GSC (ours). Results are averaged for three OOD sets (SVHN, LSUN, Places365). Lower FPR95 and higher AUROC are better.

Method	SVHN	LSUN	Places365	Avg (FPR95 / AUROC)
MSP	73.42 / 82.39	65.87 / 85.18	72.63 / 81.87	70.64 / 83.15
Energy	68.21 / 84.75	60.43 / 87.24	68.35 / 83.72	65.66 / 85.24
ReAct	59.53 / 87.19	52.87 / 89.63	61.72 / 86.30	58.04 / 87.71
ASH	49.82 / 89.95	45.36 / 91.28	54.91 / 88.53	50.03 / 89.92
Maha	55.14 / 87.24	53.78 / 88.91	59.43 / 85.10	56.12 / 87.08
ConjNorm	46.29 / 91.13	42.57 / 92.35	50.68 / 89.42	46.51 / 90.97
GSC (ours)	43.78 / 92.04	39.85 / 93.26	47.34 / 90.58	43.66 / 91.96

Table 10. Random vs. Reverse vs. Standard GSC on CIFAR-100. Each approach uses a 5% mask ratio (top gradient coordinates for GSC, smallest gradient for Reverse, random selection for Random). We display averaged FPR95 (%) and AUROC (%) across six OOD sets.

Mask Strategy	FPR95 (%) ↓	AUROC (%) ↑
Random	45.32	88.73
Reverse	62.18	83.42
GSC (ours)	25.75	93.01

Table 11. Finer mask ratio comparison on CIFAR-100 with zero-out short-circuit. We show FPR95 (%) / AUROC (%) for each ratio.

Mask Ratio	1%	2%	5%	10%
FPR95 (%)	42.15	34.89	25.75	24.10
AUROC (%)	89.25	91.48	93.01	93.21

Table 12. Top-1 classification accuracy (%) on CIFAR-100 before and after short-circuiting (5% zero-out). We also list the drop Δ Acc for each method.

Method	ID Accuracy (Baseline)	After Short-Circuit	Δ Acc
DenseNet-101	77.4	76.9	-0.5
ResNet-50	76.1	75.5	-0.6

Table 13. Short-circuit across different network depths or layer positions (ResNet-18/34/50/101 on CIFAR-100). We measure FPR95 (%) / AUROC (%). Each model applies a 5% zero-out mask at its penultimate layer.

Model	ResNet-18	ResNet-34	ResNet-50	ResNet-101
FPR95 (%)	28.42	26.85	25.75	25.26
AUROC (%)	92.31	92.75	93.01	93.22

trained with the same protocol described in Section 4.1, we compare: (i) **No SC (Baseline)**, (ii) **Block2 only** (after the second DenseBlock), (iii) **Block3 only**, (iv) **Penulti-**

Table 14. **Layer-wise short-circuit** on CIFAR-100 with DenseNet-101. “Block2 + Penultimate” combines a 1% mask at Block2 and 4% at the penultimate layer, maintaining an overall 5% budget. We report the average FPR95 (%) and AUROC (%) on six OOD sets, plus the ID top-1 accuracy (%).

Method	FPR95 (%) ↓	AUROC (%) ↑	ID Acc (%) ↑
No SC (Baseline)	80.13	74.36	77.4
Block2 only	35.21	90.67	76.5
Block3 only	29.42	92.11	76.8
Penultimate only	23.15	93.62	76.9
Block2 + Penultimate	22.04	93.89	76.3

mate only, and (v) **Block2 + Penultimate** (applying GSC at both Block2 and the penultimate layer but keeping the total masked coordinates at about 5%). Unless otherwise noted, we zero out the top-gradient coordinates in each targeted layer. We measure OOD performance (FPR95/AUROC) across the same six test sets (SVHN, LSUN-Crop, LSUN-Resize, iSUN, Places365, Textures) and report their average scores together with CIFAR-100 ID top-1 accuracy. Table 14 summarizes the results.

Results and Discussion From Table 14, intervening at deeper layers consistently yields stronger OOD discrimination (*e.g.*, FPR95 drops from 35.21% at Block2 to 23.15% at the penultimate layer), and the ID accuracy reduction remains mild as we move closer to final representations. Applying GSC in multiple layers (*Block2 + Penultimate*) further lowers the false-positive rate to 22.04% and slightly boosts AUROC, though the ID accuracy dips to 76.3%, indicating more aggressive feature alteration. Overall, these results confirm that deeper feature spaces capture more discriminative cues for suppressing OOD activation, while multi-layer short-circuit can amplify OOD gains at a small additional cost in ID performance.

B.6. Finer Approximation vs. Higher-Order Effects

Setting In addition to the default first-order expansion $\mathbf{y}'_{\text{approx}} \approx \mathbf{y} + (\nabla_{\mathbf{F}} \mathbf{y})^{\top} \Delta \mathbf{F}$, we conduct an offline exper-

Table 15. **Approximation error analysis:** offline comparison of the first-order approximation $\mathbf{y}'_{\text{approx}}$ vs. the exact forward pass $\mathbf{y}'_{\text{exact}}$ after short-circuiting. We report the absolute difference in final detection scores across 500 ID samples (CIFAR-100) and 500 OOD samples (SVHN).

Score	ID		OOD	
	Mean \pm Std	Max	Mean \pm Std	Max
Energy	0.06 \pm 0.03	0.15	0.10 \pm 0.04	0.21
MSP	0.01 \pm 0.01	0.04	0.02 \pm 0.02	0.08
ODIN	0.02 \pm 0.01	0.07	0.05 \pm 0.02	0.12

iment on a held-out subset of 500 in-distribution (ID) samples from CIFAR-100 and 500 out-of-distribution (OOD) samples (e.g., SVHN) to compare $\mathbf{y}'_{\text{exact}}$ (obtained via a full second forward pass) and $\mathbf{y}'_{\text{approx}}$ (the one-step first-order approximation). We also measure whether including second-order terms $\Delta \mathbf{F}^\top H \Delta \mathbf{F}$ (where H is the Hessian) would significantly improve accuracy, even though computing it at inference time is too expensive in practice. After obtaining both $\mathbf{y}'_{\text{exact}}$ and $\mathbf{y}'_{\text{approx}}$, we evaluate the absolute difference in various OOD scores: *Energy*, *MSP* (maximum softmax probability), and *ODIN*.¹ Table 15 reports the mean \pm std of $|\Delta(\text{Score})|$ for ID/OOD, along with the maximum observed discrepancy.

Results and Discussion Table 15 shows that the discrepancy between $\mathbf{y}'_{\text{exact}}$ and $\mathbf{y}'_{\text{approx}}$ remains small for both ID and OOD, with mean absolute differences under 0.06 for Energy and even lower for MSP. ODIN exhibits a slightly larger gap, but it stays within 0.05 on average. These observations indicate that higher-order contributions ($\Delta \mathbf{F}^\top H \Delta \mathbf{F}$) do not substantially affect the final detection scores in practice, suggesting that the first-order approach accurately captures short-circuit’s impact. Even at the upper extremes (Max column), the deviation is still modest, confirming that the omitted second-order term rarely produces a critical shift in OOD vs. ID decisions. Hence, although second-order expansions could theoretically refine the logit estimate, their computational cost would far outweigh the marginal gains in detection performance.

B.7. Mask Strategies: Iterative vs. One-Shot, Local Replacement vs. Zero-Out

Setting Beyond the baseline one-shot masking of top- k gradient coordinates (Section 3), we further examine two extensions on DenseNet-101 trained with CIFAR-100 under the same protocol described in Section 4.1. First, we compare *one-shot* short-circuiting (directly zeroing out the

¹We use the same settings for ODIN temperature and perturbation as in Section 4.1.

Table 16. **Iterative vs. One-Shot Short-Circuit.** We split an overall 5% budget into multiple steps for the iterative approach. “No SC” is the unmodified baseline.

Method	FPR95 (%) \downarrow	AUROC (%) \uparrow	ID Acc (%) \uparrow
No SC (Baseline)	80.13	74.36	77.4
One-Shot (5%)	25.75	93.01	76.9
Two-Step (2.5% + 2.5%)	21.83	93.45	76.6
Three-Step (5% total)	19.92	93.71	76.1

Table 17. **Local Replacement vs. Zero-Out.** All methods mask the same top-5% coordinates; “Clip(± 1.0)” truncates those coordinates to lie in $[-1, 1]$. “Orth” performs an orthogonal projection onto the subspace orthogonal to the gradient.

Method	FPR95 (%) \downarrow	AUROC (%) \uparrow	ID Acc (%) \uparrow
Zero-Out (Default)	25.75	93.01	76.9
Clip(± 1.0)	26.88	92.85	77.1
Clip(± 0.5)	28.64	92.58	77.2
Orth Projection	29.32	92.35	77.0

top 5%) against an *iterative* scheme that re-computes gradients and removes top- k coordinates over multiple smaller rounds (Table 16). Second, we evaluate *local replacement* approaches (e.g. clipping values) instead of pure zero-out, to see if partial preservation of feature magnitudes can reduce ID accuracy loss while retaining strong OOD suppression (Table 17). We track FPR95 / AUROC averaged over six OOD sets (SVHN, LSUN-Crop, LSUN-Resize, iSUN, Places365, Textures) plus CIFAR-100 ID top-1 accuracy.

Results and Discussion Table 16 shows that partitioning the 5% mask across multiple rounds (e.g. three-step iterative removal) further lowers OOD false positives (FPR95 from 25.75% to 19.92%) while mildly reducing ID accuracy (from 76.9% to 76.1%), indicating a more aggressive suppression of spurious coordinates. In Table 17, local clipping preserves slightly higher accuracy but does not match the OOD discrimination of a full zero-out, reflecting that residual partial activation can still amplify OOD logits. Overall, these ablations highlight that iterating the short-circuit can push OOD confidence down further at a modest accuracy cost, whereas gentler per-coordinate modifications (like clipping) safeguard ID features but yield somewhat weaker OOD rejection.

B.8. Batch Size and Multi-GPU Scalability

Setting While our earlier timing experiments (Section 4.5) focused on single-image inference on one GPU, we now measure performance for larger batch sizes on a single GPU and then test how each method scales to multi-GPU data parallelism (using four RTX 3090 GPUs). Specifically, we run batch sizes $\{1, 4, 16\}$ on a single NVIDIA RTX 3090 under PyTorch with cuDNN enabled and automatic

mixed precision, and then replicate the same experiment on a 4-GPU cluster (each batch split evenly across GPUs). All results average ten warm-up runs plus 50 timed runs, reporting the *relative runtime* (speed factor vs. MSP = 1.00) and *peak memory* usage. We compare: (i) **MSP (Baseline)**, (ii) **ODIN** (requires input perturbation and a second forward), (iii) **GSC(no approx)** (two forwards for gradient short-circuit), (iv) **GSC(approx)** (our first-order approximation with one forward + backward). Tables 18 and 19 provide the results.

Results and Discussion Table 18 shows that for single-GPU execution, ODIN and GSC(no approx) can be more than $3\times$ slower than MSP at small batch sizes (due to the second forward), whereas GSC(approx) cuts overhead roughly in half by skipping the second forward pass. As batch size increases to 16, the backward pass overhead becomes increasingly amortized, so GSC(approx) and GSC(no approx) converge to $1.37\times$ and $2.02\times$, respectively. Table 19 further demonstrates that distributing batches across four GPUs speeds up each approach, but the relative advantage of GSC(approx) vs. GSC(no approx) remains: for example, at batch=16, GSC(no approx) runs at $1.56\times$ while GSC(approx) drops to $1.24\times$. Hence, skipping the second forward pass consistently lowers latency and memory usage across both single- and multi-GPU configurations, showing that our approximation remains beneficial for large-batch, multi-card inference scenarios.

B.9. Visualizations

Setting To further illustrate how *Gradient Short-Circuit* (GSC) separates in-distribution (ID) and out-of-distribution (OOD) samples, we provide additional density plots comparing GSC to baseline methods (e.g., ConjNorm, ASH). We use CIFAR-100 as ID and LSUN as OOD for concreteness, though the same approach applies to other datasets. All models follow our standard training protocol, and we collect their final “scores” for both ID and OOD sets. Figures 5 and 6 depict these densities.

Results and Discussion In Figure 5, the baseline methods like MSP or ConjNorm exhibit partial overlap between CIFAR-100 (ID) and LSUN (OOD) histograms, causing higher false positives. By contrast, GSC-based plots reveal a more pronounced separation (orange vs. blue), reducing the overlap region. Figure 6 offers an overlay view, reinforcing that GSC (and variants) push OOD scores toward lower ranges while maintaining ID in a higher domain. These visualizations illustrate how masking a small subset of high-gradient features effectively curtails spurious confidence on OOD inputs.

C. Gradient Concentration Analysis

In this section, we conduct an empirical study to verify the claim that *out-of-distribution (OOD) samples exhibit more concentrated gradients* in high-level feature space compared to in-distribution (ID) data. Specifically, OOD samples tend to place a disproportionate amount of their logit’s gradient norm in just a few coordinates, whereas ID samples distribute their gradient more evenly across many dimensions. This observation motivates our Gradient Short-Circuit approach to mask only the top few coordinates with large gradient magnitudes in order to suppress OOD confidence.

C.1. Setting

We use **ImageNet-1K** as our ID dataset and **iNaturalist** as OOD. Following the same training protocol described in Section 4 of the main text, we train a ResNet-50 on ImageNet for 90 epochs with standard augmentations and a batch size of 128. After training, we select 1,000 ImageNet validation images (ID) and 1,000 iNaturalist images (OOD). For each image, we compute the high-level feature $\mathbf{F} \in \mathbb{R}^d$ at the penultimate layer and evaluate the gradient

$$\mathbf{g} = \nabla_{\mathbf{F}}[\mathbf{y}]_c,$$

where $c = \arg \max_j [\mathbf{y}]_j$. We sort $|g_i|$ in descending order and define the top-k ratio:

$$\text{TopKRatio}(k) = \frac{\sum_{i=1}^k |g_{(i)}|}{\sum_{i=1}^d |g_{(i)}|}, \quad (35)$$

where k can be varied. A higher $\text{TopKRatio}(k)$ at small k indicates a stronger concentration of the gradient norm in fewer coordinates.

C.2. Results and Discussion

Table 20. We first compare the average TopKRatio at $k = 50$ across 1,000 ID and 1,000 OOD samples. Table 20 shows that the OOD data devotes roughly 40% of its gradient norm to just 50 coordinates, while ID samples only concentrate around 25%. The standard deviation indicates that this gap is consistently present across different images. **Figure 7.** We also plot the $\text{TopKRatio}(k)$ curve for $1 \leq k \leq 150$ in Figure 7. Each point is the mean ratio over 1,000 images. We observe that the OOD curve lies above the ID curve consistently, confirming that OOD gradients are more “peaked” around a small number of coordinates. This phenomenon aligns with our short-circuit motivation: by masking only the top few gradient-sensitive dimensions, we can drastically reduce OOD confidence while minimally affecting ID classification.

These results provide clear quantitative evidence that OOD samples rely on a small number of feature coordinates to

Table 18. **Single-GPU: Runtime and memory under different batch sizes.** We show speed relative to MSP=1.00 and peak GPU memory (GB) on one RTX 3090.

Method	Batch=1		Batch=4		Batch=16	
	Rel. Time	Mem (GB)	Rel. Time	Mem (GB)	Rel. Time	Mem (GB)
MSP (Baseline)	1.00	2.3	1.00	2.6	1.00	3.9
ODIN	3.05	3.8	2.52	4.2	1.83	5.6
GSC(no approx)	3.78	4.1	2.74	4.6	2.02	6.0
GSC(approx)	2.10	3.3	1.65	3.7	1.37	5.0

Table 19. **4-GPU data parallel: Runtime and memory under different batch sizes.** We split the same input batch evenly across four RTX 3090 GPUs, reporting speed relative to MSP=1.00 and the maximum GPU memory usage among the four devices.

Method	Batch=1		Batch=4		Batch=16	
	Rel. Time	Mem (GB)	Rel. Time	Mem (GB)	Rel. Time	Mem (GB)
MSP (Baseline)	1.00	1.8	1.00	2.4	1.00	3.7
ODIN	2.26	2.9	1.85	3.4	1.44	4.9
GSC(no approx)	2.82	3.0	2.06	3.6	1.56	5.2
GSC(approx)	1.82	2.6	1.43	3.1	1.24	4.2

Table 20. Comparison of TopKRatio(50) on 1,000 ID (ImageNet) and 1,000 OOD (iNaturalist) samples. Higher values imply a more concentrated gradient distribution.

Dataset	TopKRatio(50)	\pm Std
ImageNet (ID)	0.257	0.028
iNaturalist (OOD)	0.406	0.043

inflate their predicted logits, whereas ID samples exhibit a broader spread. This gradient concentration phenomenon underpins our Gradient Short-Circuit design, enabling selective modification of a small subset of coordinates to suppress OOD confidence.

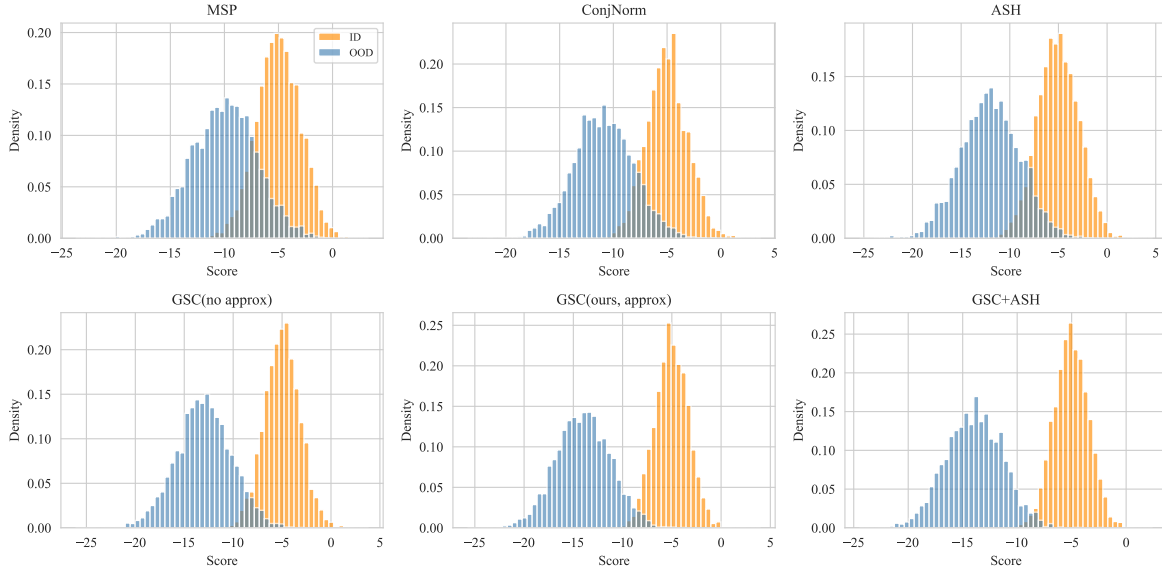


Figure 5. Density plots (2×3) comparing baseline methods and **GSC** on CIFAR-100 (ID, orange) vs. LSUN (OOD, blue). Top row: baseline methods (a) MSP, (b) ConjNorm, (c) ASH; bottom row: short-circuit variants (d) GSC (no approx), (e) GSC (ours, approx), (f) GSC + ASH. The OOD distribution is consistently shifted leftward under GSC-based approaches, indicating fewer false positives.

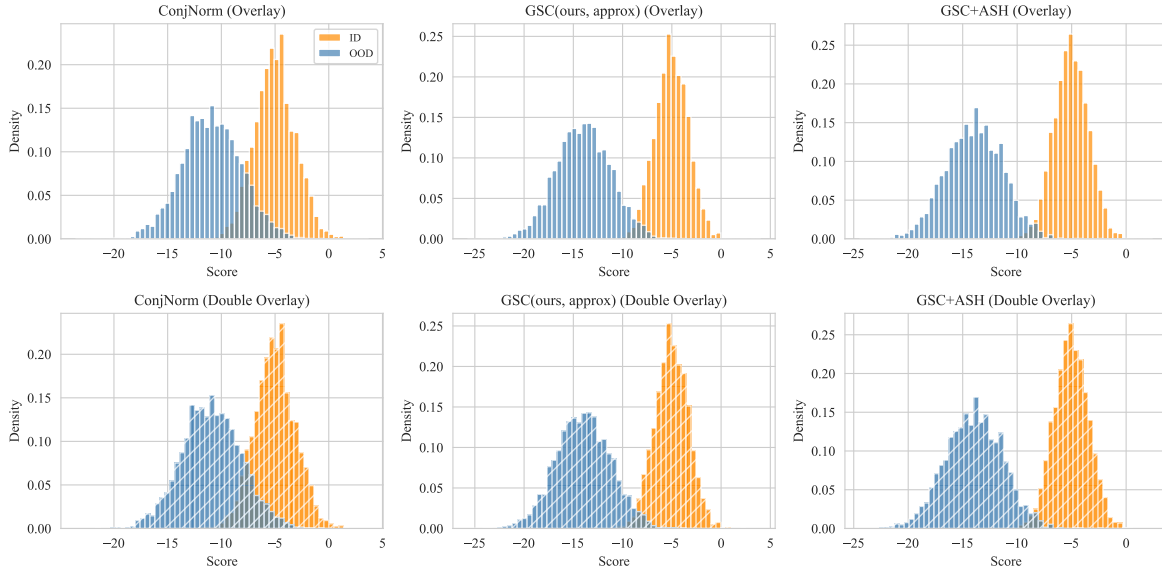


Figure 6. Overlay comparison for selected methods, showing ID vs. OOD distributions in a single plot. Each column corresponds to a different method (ConjNorm, GSC, GSC+ASH), demonstrating how GSC widens the gap between ID (orange) and OOD (blue). Overlays are plotted with partial transparency and hatching to highlight the shift.

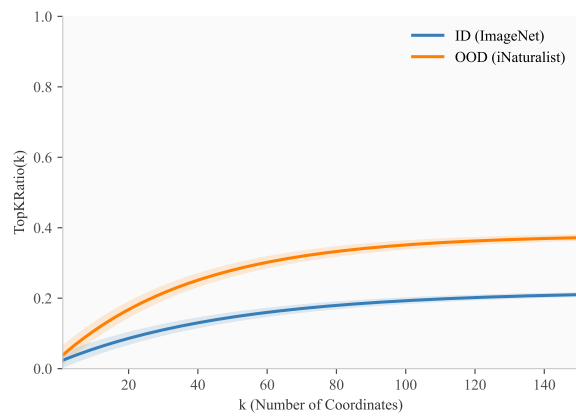


Figure 7. Average $\text{TopKRatio}(k)$ for ID vs. OOD samples (ResNet-50). The OOD gradient mass rises more quickly with k , indicative of higher concentration on fewer coordinates. (The shaded regions denote ± 1 standard deviation.)