

## Appendix

<b>1</b>	<b>Theoretical Analysis: LogitsGap vs Softmax Scaling</b>	<b>2</b>
<b>2</b>	<b>Analysis: Logits Distribution Differences Between ID and OOD Samples</b>	<b>5</b>
2.1	Motivation Example . . . . .	5
2.2	Motivation Evidence . . . . .	5
2.3	Theoretical Analysis . . . . .	5
<b>3</b>	<b>Experimental Details</b>	<b>7</b>
3.1	Implementation Details . . . . .	7
3.2	Datasets . . . . .	7
3.3	Logits Selection Methods . . . . .	8
3.4	Implementation Details on LogitGap Combined with Negative-Prompt-Based OOD Method . . . . .	8
<b>4</b>	<b>Experimental Results</b>	<b>9</b>
4.1	Results on Traditional OOD Detection with CIFAR100 as ID Dataset . . . . .	9
4.2	Results on Traditional OOD Detection with ImageNet as ID Dataset . . . . .	9
4.3	Results on Far-OOD Evaluation with ImageNet as ID Dataset in Zero-Shot OOD Detection . . . . .	10
4.4	Results on More Architectures with ImageNet as ID Dataset in Zero-Shot OOD Detection . . . . .	10
<b>5</b>	<b>More Analysis about LogitGap</b>	<b>12</b>
5.1	Generalization Beyond Visual Tasks . . . . .	12
5.2	Relationship Between LogitGap and Other Logit-Pattern-Based Methods . . . . .	12
5.3	The Effectiveness of Synthetic OOD Data . . . . .	13
5.4	The Impact of Hyperparameter N on LogitGap Performance . . . . .	13

## 1 Theoretical Analysis: LogitsGap vs Softmax Scaling

In this section, we provide the proof of Theorem 4.1 from Section 4, which offers a detailed analysis of the relationship between LogitGap and MCM [23]. We first introduce the necessary notation and definitions.

**Notations.** Formally, in the out-of-distribution problem, the decision function  $D : \mathcal{X} \rightarrow \{\text{ID}, \text{OOD}\}$  is typically constructed as:

$$D(\mathbf{x}) = \begin{cases} \text{ID}, & \text{if } S(\mathbf{x}; f) \geq \lambda \\ \text{OOD}, & \text{if } S(\mathbf{x}; f) < \lambda \end{cases} \quad (1)$$

where  $S : \mathcal{X} \rightarrow \mathbb{R}$  is a scoring function based on pre-trained model  $f$ , which assigns a scalar confidence score to each input sample. The decision threshold  $\lambda$  establishes the decision boundary: samples satisfying  $S(\mathbf{x}) < \lambda$  are identified as OOD, while those with  $S(\mathbf{x}) \geq \lambda$  are considered ID.

We present the OOD scoring functions for LogitGap and MCM. Given a logit vector  $\mathbf{z}$  predicted for a test sample  $\mathbf{x}$ , we sort its elements in descending order to obtain  $\mathbf{z}'$ . LogitGap exploits the inherent distributional differences between ID and OOD logits<sup>1</sup>:

$$S_{\text{LogitGap}}(\mathbf{x}; f) = \frac{1}{K} \sum_{j=1}^K (z'_1 - z'_j), \quad (2)$$

where  $z'_j$  denotes the  $j$ -th largest logit.

MCM [23] applies softmax normalization to the logits and takes the maximum probability as the OOD score:

$$S_{\text{MCM}}(\mathbf{x}; f) = \max_k \frac{e^{z_k/\tau}}{\sum_{j=1}^K e^{z_j/\tau}}, \quad (3)$$

where  $\tau$  is a temperature scaling parameter.

Furthermore, we give the OOD decision function based on the corresponding score:

$$D_{\text{LogitGap}}(\mathbf{x}) = \begin{cases} \text{ID}, & \text{if } S_{\text{LogitGap}}(\mathbf{x}; f) \geq \lambda_{\text{LogitGap}} \\ \text{OOD}, & \text{if } S_{\text{LogitGap}}(\mathbf{x}; f) < \lambda_{\text{LogitGap}} \end{cases}, \quad (4)$$

$$D_{\text{MCM}}(\mathbf{x}) = \begin{cases} \text{ID}, & \text{if } S_{\text{MCM}}(\mathbf{x}; f) \geq \lambda_{\text{MCM}} \\ \text{OOD}, & \text{if } S_{\text{MCM}}(\mathbf{x}; f) < \lambda_{\text{MCM}} \end{cases}. \quad (5)$$

**Remarks.** By convention,  $\lambda_{\text{LogitGap}}$  and  $\lambda_{\text{MCM}}$  are typically chosen such that the true positive rate is at 95%. For brevity, we will use  $\lambda_L$  to denote  $\lambda_{\text{LogitGap}}$  throughout the rest of this paper.

**Theorem 1.1.** *Given a  $K$ -way classification task, let the predicted logit vector for a sample  $\mathbf{x}$  be  $\mathbf{z} = [z_1, z_2, \dots, z_K]$ . Let  $\mathbf{z}' = [z'_1, z'_2, \dots, z'_K]$  denote the sorted logits in descending order, such that  $z'_1 = \max_k z_k$ . Then, for the temperature scaling parameter  $\tau$  used in the softmax function, if  $\tau > 2(K-1)$ , we have*

$$\text{FPR}_{\text{LogitGap}}(\lambda_L) \leq \text{FPR}_{\text{MCM}}(\tau, \lambda_{\text{MCM}}),$$

where  $\text{FPR}_{\text{LogitGap}}(\lambda_L)$  is the false positive rate based on LogitGap with threshold  $\lambda_L$ . Similarly,  $\text{FPR}_{\text{MCM}}(\tau, \lambda_{\text{MCM}})$  is the false positive rate based on MCM with temperature  $\tau$  and threshold  $\lambda_{\text{MCM}}$ .

*Proof.* Let  $Q_x$  denotes the out-of-distribution  $P_{\mathbf{x}|\text{OOD}}$ . By definition, we express the false positive rate of MCM, denoted as  $\text{FPR}_{\text{MCM}}(\tau, \lambda_{\text{MCM}})$ , in the following:

$$\begin{aligned} \text{FPR}_{\text{MCM}}(\tau, \lambda_{\text{MCM}}) &= Q_x(S_{\text{MCM}}(\mathbf{x}; f) > \lambda_{\text{MCM}}) \\ &= Q_x\left(\frac{e^{z'_1/\tau}}{\sum_{j=1}^K e^{z'_j/\tau}} > \lambda_{\text{MCM}}\right). \end{aligned} \quad (6)$$

---

<sup>1</sup>For ease of derivation, we adopt Equation 2 in place of  $S_{\text{LogitGap}}(\mathbf{x}; f) = \frac{1}{K-1} \sum_{j=2}^K (z'_1 - z'_j)$ , without affecting performance.

Next, we introduce a intermediate OOD score function, LogitGap with softmax (LogitGap\_softmax) to bridge the MCM and our LogitGap. We define the score function of LogitGap\_softmax as  $S_{\text{LogitGap\_softmax}}(\mathbf{x}; f) = \frac{\sum_{i=1}^K [e^{z'_1/\tau} - e^{z'_i/\tau}]}{K \sum_{j=1}^K e^{z'_j/\tau}}$ . Therefore, the false positive rate of LogitGap\_softmax can be expressed as

$$\begin{aligned} \text{FPR}_{\text{LogitGap\_softmax}}(\tau, \lambda_{\text{LM}}) &= Q_x (S_{\text{LogitGap\_softmax}}(\mathbf{x}; f) > \lambda_{\text{LM}}) \\ &= Q_x \left( \frac{\sum_{i=1}^K [e^{z'_1/\tau} - e^{z'_i/\tau}]}{K \sum_{j=1}^K e^{z'_j/\tau}} > \lambda_{\text{LM}} \right) \\ &= Q_x \left( \frac{e^{z'_1/\tau}}{\sum_{j=1}^K e^{z'_j/\tau}} - \frac{1}{K} > \lambda_{\text{LM}} \right) \\ &= Q_x \left( \frac{e^{z'_1/\tau}}{\sum_{j=1}^K e^{z'_j/\tau}} > \lambda_{\text{LM}} + \frac{1}{K} \right) \end{aligned} \quad (7)$$

Combining Eq.(6) and Rq.(7), we have  $\lambda_{\text{MCM}} = \lambda_{\text{LM}} + \frac{1}{K}$ . Similarly, we can express the false positive rate of LogitGap as

$$\begin{aligned} \text{FPR}_{\text{LogitGap}}(\lambda_{\text{L}}) &= Q_x (S_{\text{LogitGap}}(\mathbf{x}; f) > \lambda_{\text{L}}) \\ &= Q_x \left( \frac{\sum_{j=1}^K (z'_1 - z'_j)}{K} > \lambda_{\text{L}} \right) \\ &= Q_x \left( \frac{\sum_{j=1}^K (z'_1 - z'_j)}{\tau * K} > \frac{1}{\tau} * \lambda_{\text{L}} \right) \end{aligned}$$

Since the outputs of CLIP are bounded within  $[-1, 1]$ <sup>2</sup>, it follows that

$$\sum_{j=1}^K (z'_1 - z'_j) \leq 2(K-1).$$

Furthermore, we have

$$\begin{aligned} \text{FPR}_{\text{LogitGap}}(\lambda_{\text{L}}) &= Q_x \left( \frac{\sum_{j=1}^K (z'_1 - z'_j)}{\tau * K} > \frac{1}{\tau} * \lambda_{\text{L}} \right) \\ &\leq Q_x \left( \frac{2(K-1)}{\tau * K} > \frac{1}{\tau} * \lambda_{\text{L}} \right) \\ &= Q_x \left( \frac{2(K-1)}{\tau} * \frac{1}{K} > \frac{1}{\tau} * \lambda_{\text{L}} \right) \\ &\leq Q_x \left( \frac{2(K-1)}{\tau} * \frac{e^{z'_1/\tau}}{\sum_{j=1}^K e^{z'_j/\tau}} > \frac{1}{\tau} * \lambda_{\text{L}} \right) \\ &= Q_x \left( \frac{e^{z'_1/\tau}}{\sum_{j=1}^K e^{z'_j/\tau}} > \frac{\lambda_{\text{L}}}{2(K-1)} \right) \end{aligned} \quad (8)$$

Next, for  $\lambda_{\text{LM}}$  and  $\lambda_{\text{L}}$ , we consider two cases: (1)  $\lambda_{\text{LM}} \leq \frac{\lambda_{\text{L}}}{\tau}$ ; (2)  $\lambda_{\text{LM}} > \frac{\lambda_{\text{L}}}{\tau}$ .

For the case (1), with the assumption  $\lambda_{\text{LM}} \leq \frac{\lambda_{\text{L}}}{\tau}$ , we have

$$\text{FPR}_{\text{LogitGap}}(\lambda_{\text{L}}) \leq Q_x \left( \frac{e^{z'_1/\tau}}{\sum_{j=1}^K e^{z'_j/\tau}} > \frac{\tau * \lambda_{\text{LM}}}{2(K-1)} \right) \quad (9)$$

---

<sup>2</sup>For non-CLIP models, assuming the output range is  $[-A, A]$ , we similarly obtain  $\sum_{j=1}^K (z'_1 - z'_j) \leq 2A(K-1)$

By comparing Eq.(6) and Eq.(9), we can derive that  $\text{FPR}_{\text{LogitGap}}(\lambda_L) \leq \text{FPR}_{\text{MCM}}(\tau, \lambda_{\text{MCM}})$ , if the following condition holds:

$$\frac{\tau * \lambda_{\text{LM}}}{2(K-1)} \geq \lambda_{\text{MCM}}. \quad (10)$$

Note that,  $\lambda_{\text{LM}} = \lambda_{\text{MCM}} - \frac{1}{K}$ . Combining this equation with Eq.(10), we have

$$\tau \geq 2(K-1) * \frac{\lambda_{\text{MCM}}}{\lambda_{\text{LM}}} > 2(K-1) \quad (11)$$

For the case (2), by directly comparing Eq.(6) and Eq.(8), we can derive that  $\text{FPR}_{\text{LogitGap}}(\lambda_L) \leq \text{FPR}_{\text{MCM}}(\tau, \lambda_{\text{MCM}})$ , if the following condition holds:

$$\frac{\lambda_L}{2(K-1)} \geq \lambda_{\text{MCM}}. \quad (12)$$

In the case (2), we have  $\lambda_{\text{LM}} > \frac{\lambda_L}{\tau}$ . With this inequation, Eq.(12) can be rewritten as

$$\frac{\tau * \lambda_{\text{LM}}}{2(K-1)} > \frac{\lambda_L}{2(K-1)} \geq \lambda_{\text{MCM}}. \quad (13)$$

Note that Eq.(13) is the same as Eq.(11), thus we can derive the same conclusion.  $\square$

## 2 Analysis: Logits Distribution Differences Between ID and OOD Samples

### 2.1 Motivation Example

To illustrate a key limitation of the softmax-based scoring function MCM [23], we present two examples where *logit distributions differ significantly yet yield nearly identical MCM scores*. This reveals how the softmax operation can suppress informative structural differences in the logit space, which are often critical for distinguishing ID and OOD samples. Consider a 3-way classification problem with two test samples,  $\mathbf{x}^1$  and  $\mathbf{x}^2$ , whose corresponding logit vectors are  $\mathbf{z}^1 = [0.5596, -0.9808, -0.9808]$  and  $\mathbf{z}^2 = [0.9783, -0.6311, -0.4976]$ , respectively. Despite the significant difference in magnitude and sharpness ( $\mathbf{z}^2$  exhibits a much more confident prediction), the maximum softmax probability for both is identical:  $\text{softmax}(\mathbf{z}^1) = [0.70, 0.15, 0.15]$  and  $\text{softmax}(\mathbf{z}^2) = [0.70, 0.14, 0.16]$ .

This arises because softmax normalizes logits into a probability distribution, preserving their relative ordering while discarding absolute magnitudes. As a result, distinct logit patterns can be mapped to probability vectors with similar maximum values, thereby limiting the discriminative power of softmax-based OOD detection. This issue is further exacerbated by temperature scaling  $\tau > 1$ , which flattens the softmax distribution and amplifies information loss. This observation naturally raise a key question: *How can we more effectively leverage the discriminative information embedded in non-maximum logits to enhance ID-OOD separability?*

### 2.2 Motivation Evidence

We observe an interesting phenomenon: the relationship between the maximum logit and the remaining logits differs notably between in-distribution (ID) and out-of-distribution (OOD) samples. As shown in Figure 1, ID samples tend to have a larger maximum logit value than OOD samples. Conversely, for the non-maximum logits, OOD samples typically exhibit higher values than ID samples.

To verify the generality of this phenomenon, we conduct experiments across various ID/OOD datasets and model architectures in Table 1. Specifically, we report the average maximum logit, and the average logit of non-predicted classes, using ResNet-50, ViT-B/16, and ViT-L/14 as backbone models.

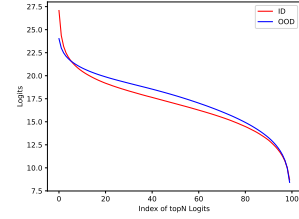


Figure 1: Descending-sorted logits on CLIP ViT-B/16 with ImageNet100 (ID) and iNaturalist (OOD).

### 2.3 Theoretical Analysis

Our theoretical motivation stems from the observed distinction in logit patterns between ID and OOD samples: (i) Higher maximum logit values for ID samples; (ii) Higher non-predicted class logits in OOD samples. This indicates that the gap between the maximum logit and the remaining logits is generally wider for ID samples than for OOD samples.

To understand the rationale behind this, we provide a simple analysis from the theoretical perspective.

**Theorem 2.1.** *In a binary classification problem, given a well-trained feature extractor  $\phi$  and a classifier  $W$ , we assume that the feature distribution of ID samples for class  $i$  follows a Gaussian distribution,  $\phi(\mathbf{x}_{\text{ID}}|y = 0) \sim \mathcal{N}(\mu_0, \Sigma_0)$  and  $\phi(\mathbf{x}_{\text{ID}}|y = 1) \sim \mathcal{N}(\mu_1, \Sigma_1)$ . We further assume that the features of OOD samples can be modeled as an interpolation of two ID feature distribution with a Gaussian noise, i.e.  $\phi(\mathbf{x}_{\text{OOD}}) = \alpha \cdot \mathcal{N}(\mu_1, \Sigma_1) + (1 - \alpha) \cdot \mathcal{N}(\mu_0, \Sigma_0) + \beta \cdot \mathcal{N}(0, \Sigma)$ . Let  $\mathbf{z}^{\text{ID}} = [z_0^{\text{ID}}, z_1^{\text{ID}}]$  denotes the predicted logit vector for ID sample  $\mathbf{x}_{\text{ID}}$ ,  $\mathbf{z}^{\text{OOD}} = [z_0^{\text{OOD}}, z_1^{\text{OOD}}]$  denotes the predicted logit vector for OOD sample  $\mathbf{x}_{\text{OOD}}$ . If the ID sample  $\mathbf{x}_{\text{ID}}$  belongs to class  $i$ , we have*

$$\mathbb{E}[z_{1-i}^{\text{ID}}] < \mathbb{E}[z_{1-i}^{\text{OOD}}] \quad (14)$$

*Proof.* If the ID sample  $\mathbf{x}_{\text{ID}}$  belongs to class  $i$ , we have

$$\mathbb{E}[z_{1-i}^{\text{ID}}] = w_{1-i}^T \mathbb{E}[\phi(\mathbf{x}_{\text{ID}})] = w_{1-i}^T \mu_i \quad (15)$$

Table 1: Logit distribution differences between ID and OOD data.  $z'_1$  and  $\bar{z}'_N$  denote the maximum logit and the average logit across all non-predicted classes, respectively.

	$z'_1$	$\bar{z}'_N$
CIFAR-10 (ID)	9.02	-0.12
CIFAR-100	7.69	0.40
TIN	7.19	0.57
Textures	5.54	0.85
SVHN	5.51	0.84

(a) Logit distribution on ResNet-50 using CIFAR-10 as ID dataset.

	$z'_1$	$\bar{z}'_N$
ImageNet (ID)	30.82	17.32
iNaturalist	26.48	17.89
SUN	26.71	17.90
Textures	28.72	18.94

(b) Logit distribution on CLIP ViT-B/16 using ImageNet as ID dataset.

	$z'_1$	$\bar{z}'_N$
ImageNet (ID)	25.76	11.19
iNaturalist	20.72	11.65
Textures	23.85	13.18
ImageNet-O	21.14	12.16

(c) Logit distribution on CLIP ViT-L/14 using ImageNet as ID dataset.

Similarly, we have

$$\mathbb{E}[z_{1-i}^{\text{OOD}}] = w_{1-i}^T \mathbb{E}[\phi(\mathbf{x}_{\text{OOD}})]. \quad (16)$$

Assuming the features of  $\mathbf{x}_{\text{OOD}}$  can be denoted as the interpolation of two ID feature distribution with a random Gaussian noise, *i.e.*  $\phi(\mathbf{x}_{\text{OOD}}) = \alpha \cdot \mathcal{N}(\mu_1, \Sigma_1) + (1 - \alpha) \cdot \mathcal{N}(\mu_0, \Sigma_0) + \beta \cdot \mathcal{N}(0, \Sigma)$ , we have

$$\begin{aligned} \mathbb{E}[z_{1-i}^{\text{OOD}}] &= w_{1-i}^T \mathbb{E}[\phi(\mathbf{x}_{\text{OOD}})] \\ &= w_{1-i}^T (\alpha \cdot \mu_1 + (1 - \alpha) \cdot \mu_0) \\ &= \alpha \cdot w_{1-i}^T \mu_1 + (1 - \alpha) \cdot w_{1-i}^T \mu_0. \end{aligned} \quad (17)$$

We assume that the classifier  $W$  is well trained, thus  $w_{1-i}^T \mu_i < w_{1-i}^T \mu_{1-i}$ . Then, we have

$$w_{1-i}^T \mu_i < \alpha \cdot w_{1-i}^T \mu_1 + (1 - \alpha) w_{1-i}^T \mu_0. \quad (18)$$

Therefore, we have

$$\mathbb{E}[z_{1-i}^{\text{ID}}] < \mathbb{E}[z_{1-i}^{\text{OOD}}]. \quad (19)$$

□

### 3 Experimental Details

#### 3.1 Implementation Details

We run all OOD detection experiments on NVIDIA GeForce RTX-4090Ti GPUs with Pytorch 2.3.1. For CLIP-based OOD detection, we adopt the pre-trained ViT-B/16 [8] model from CLIP [26]. For conventional OOD detection task using CIFAR-10 [18] as ID dataset, we use a ResNet-50 [10] model pre-trained on the CIFAR-10 training set as the backbone.

#### 3.2 Datasets

**ImageNet-10, ImageNet-20, ImageNet100** [23] creates ImageNet-10 that mimics the class distribution of CIFAR-10 [18] but with high-resolution images. For semantically hard OOD evaluation with realistic datasets, [23] curates ImageNet-20, which consists of 20 classes semantically similar to ImageNet-10 (e.g., dog (ID) vs. wolf (OOD)). ImageNet-100 is created by randomly sample 100 classes from ImageNet-1k [5].

**NINCO** The NINCO [2] main dataset comprises 64 OOD classes with a total of 5,879 samples. These classes were carefully selected to ensure no categorical overlap with any of the ImageNet-1K [5] classes. Additionally, each sample was manually inspected by the authors to confirm the absence of ID objects, making NINCO a reliable benchmark for evaluating out-of-distribution detection on ImageNet-1K.

**ImageNetOOD** ImageNet-OOD [37] is a clean, manually curated, and diverse dataset containing 31,807 images from 637 classes, designed for evaluating semantic shift detection using ImageNet-1K [5] as the in-distribution (ID) dataset. To minimize covariate shifts, images are sourced directly from ImageNet-21K [28], with human verification ensuring the removal of any ID contamination from ImageNet-1K. The dataset also addresses multiple sources of semantic ambiguity caused by inaccurate hierarchical relationships in ImageNet labels and eliminates visually ambiguous images stemming from inconsistencies in ImageNet’s data curation process.

**ImageNet-O** ImageNet-O is a dataset containing image concepts absent from ImageNet-1K [5], specifically designed to evaluate the robustness of ImageNet models. These out-of-distribution (OOD) images consistently cause models to misclassify them as high-confidence in-distribution examples. As the first anomaly and OOD dataset tailored for testing ImageNet-1K models, ImageNet-O provides a valuable benchmark for assessing OOD detection performance under label distribution shifts.

**ImageNet-A** ImageNet-A [16] contains images sampled from a distribution distinct from the standard ImageNet-1k [5] training distribution. Although its examples belong to existing ImageNet-1k classes, they are deliberately more challenging, frequently leading to misclassifications across a range of models. ImageNet-A enables us to test image classification performance when the input data has covariate distribution shifts.

**ImageNet-R** ImageNet-R (Rendition) [13] consists of artistic and non-photographic renditions of ImageNet-1k [5] classes, including art, cartoons, graffiti, embroidery, graphics, origami, paintings, patterns, plastic figures, plush toys, sculptures, sketches, tattoos, video game assets, and more. The dataset covers 200 ImageNet-1k classes with a total of 30,000 images, offering a diverse benchmark for evaluating model robustness to distribution shifts in visual style and appearance.

**ImageNet-Sketch** The ImageNet-Sketch [34] dataset contains 50,889 images, with approximately 50 images per class for all 1,000 ImageNet-1k [5] categories. It was constructed by querying Google Images using the phrase “sketch of [class name],” restricted to a black-and-white color scheme. An initial set of 100 images was collected for each class, followed by manual filtering to remove irrelevant images and those belonging to visually similar but incorrect classes. For categories with fewer than 50 valid images after cleaning, data augmentation techniques such as flipping and rotation were applied to balance the dataset.

Table 2: The value of  $N$  for two methods on different datasets.

	$K$	$N$ (Fixed)	$N$ (Adaptive)
ImageNet	1000	200	88
ImageNet-100	100	20	20
ImageNet-10	10	5	4
ImageNet-20	20	10	6

### 3.3 Logits Selection Methods

**Logits Selection with a Fixed Hyperparameter  $N$**  The effectiveness of our proposed LogitGap method relies on the selection of an informative region within the logits space, which is governed by a hyperparameter  $N$ . To determine an appropriate value for  $N$ , we consider the number of classes in the dataset. For datasets with a large number of classes (e.g., ImageNet and ImageNet-100), we set  $N$  to 20% of the total number of classes  $K$ . In contrast, for datasets with fewer classes (e.g., ImageNet-10 and ImageNet-20), we set  $N$  to 50% of  $K$ . This strategy is motivated by the following intuition: when  $K$  is large, using a lower proportion (e.g., 20%) helps reduce the influence of noise; whereas when  $K$  is small, a higher proportion (e.g., 50%) preserves more useful information. Moreover, our empirical results demonstrate a consistent performance improvement when  $N$  is set within the 20%–50% range of the total number of classes across different datasets. The specific  $N$  values chosen for different datasets are summarized in Table 2.

**Logits Selection with an Adaptive Hyperparameter  $N$**  To enhance LogitGap, we introduce an improved strategy to adaptively choose the value of  $N$ . To this end, we firstly construct a small validation set randomly sampled from the in-distribution (ID) data, with a fixed size of 100 samples. Using the model’s feature extractor, we obtain the image features of these ID samples. We then perform random inter-class interpolation on these features to generate synthetic OOD samples. To further increase the diversity of these synthetic samples, Gaussian noise is added to the interpolated features. The interpolation process is formally defined as follows:

$$\mathbf{x}_{\text{OOD}} = \alpha \cdot \mathbf{x}_i + (1 - \alpha) \cdot \mathbf{x}_j + \beta \cdot \mathcal{N}(0, I), \quad (20)$$

where  $\mathbf{x}_i$  and  $\mathbf{x}_j$  represent two samples from the ID validation set,  $\alpha$  denotes the mixing coefficient between samples, while  $\beta$  controls the weight of the added noise.

In our experiments, we generate a set of synthetic OOD samples equal in size to the ID validation set to serve as the OOD validation set. For datasets with a large number of categories, such as ImageNet [5] and ImageNet-100 [23], we set the interpolation parameters to  $\alpha = 0.3$  and  $\beta = 0.8$ . For smaller-scale datasets like ImageNet-10 [23] and ImageNet-20 [23], we set  $\alpha = 0.3$  and  $\beta = 0.0$ . We then compute the logits scores for both the ID and synthetic OOD validation samples and determine the optimal value of  $N$  adaptively based on the following criterion:

$$\arg \max_N (\mathbb{E}_{\mathbf{x} \sim P_{\text{OOD}}} [\bar{z}'_N] - \mathbb{E}_{\mathbf{x} \sim P_{\text{ID}}} [\bar{z}'_N]), \quad (21)$$

where  $\bar{z}'_N$  represents the mean of the logits ranked from second to the  $N$ -th largest.

### 3.4 Implementation Details on LogitGap Combined with Negative-Prompt-Based OOD Method

Negative-Prompt-Based OOD detection methods jointly train both positive and negative prompts to separately capture the characteristics of ID and OOD samples. In these methods, the model’s predicted logits take the form  $\mathbf{z} = [z_1^{\text{ID}}, z_2^{\text{ID}}, \dots, z_K^{\text{ID}}, z_1^{\text{OOD}}, z_2^{\text{OOD}}, \dots, z_M^{\text{OOD}}]$ , where  $z_i^{\text{ID}}$  denotes the logit for the  $i$ -th ID class, and  $z_j^{\text{OOD}}$  denotes the logit for the  $j$ -th OOD class. Since negative prompts tend to increase the logits of OOD samples, directly applying LogitGap over the entire logits space is not appropriate. In our experiments, when combining ID-Like [1] with LogitGap, we rely solely on the ID logits,  $\mathbf{z}^{\text{ID}} = [z_1^{\text{ID}}, z_2^{\text{ID}}, \dots, z_K^{\text{ID}}]$ , to compute the LogitGap score.



## 4 Experimental Results

### 4.1 Results on Traditional OOD Detection with CIFAR100 as ID Dataset

To further evaluate the effectiveness and generality of our LogitGap, we conduct experiments in traditional OOD detection settings with CIFAR100 [18] as ID dataset. Following the OpenOOD [36] protocol, we adopt the standard OpenOOD benchmark splits for OOD evaluation. The OOD benchmarks include both **near-OOD datasets**: CIFAR-10 [18], Tiny ImageNet [31], and **far-OOD datasets**: MNIST [6], SVHN [24], Textures [3], and Places365 [41]. In this setup, we employ a ResNet-50 [10] backbone trained from scratch on ID data using cross-entropy loss. We compare LogitGap against a comprehensive suite of baselines: (i) Logit-based post-hoc methods: MSP [14], MaxLogit [11], KL-Matching [12], ODIN [20], IODIN [27]; (ii) Internal network statistics-based methods: Mahalanobis [19], ASH [7], SHE (Hopfield energy) [39]; (iii) Training-based methods with auxiliary data: Outlier Exposure (OE) [15], MixOE [38]. The results are summarized in Table 3.

As shown in Table 3, our proposed LogitGap demonstrates superior performance across all evaluated benchmarks. We observe that: (1) Among logit-based methods, IODIN [27] achieves the best performance. ODIN [20] enhances OOD detection by introducing input perturbations to amplify the difference between ID and OOD samples in the softmax output. Building upon this, IODIN proposes to mask low-magnitude regions and perturb only invariant features, effectively encouraging the model to ignore environmental factors and focus on more informative regions. This requires calculating an additional invariant feature mask on top of the gradient computation in ODIN. In contrast, LogitGap improves detection performance by applying a simple transformation in the output space, reducing the FPR95 by 1.98% and 0.78% compared to MSP [14] and IODIN, respectively. (2) Internal network statistics-based methods generally perform worse because they heavily rely on the representation quality and discriminative ability of the internal network. When the backbone lacks sufficient expressive power, the extracted feature distributions of ID and OOD samples tend to overlap, making it difficult to effectively distinguish between them. (3) When combined with methods that leverage extra out-of-distribution data, such as OE [15] and MixOE [38], LogitGap further enhances performance over their respective baselines. For example, a straightforward substitution of the OOD score function with LogitGap leads to a 3.57% and 1.23% performance gain in FPR95 and AUROC respectively over the original OE [15] method. These experiments highlight the advantages of LogitGap in terms of ease of integration and wide applicability across different settings.

Table 3: OOD Detection performance on ResNet-50 using CIFAR-100 as the ID dataset.

Method	Near OOD								Far OOD								AVG	
	CIFAR10		TIN		MNIST		SVHN		Textures		PLACES365							
	FPR95	AUROC	FPR95	AUROC	FPR95	AUROC	FPR95	AUROC	FPR95	AUROC	FPR95	AUROC	FPR95	AUROC	FPR95 ↓	AUROC ↑		
MSP [14]	58.90	78.47	50.70	82.07	57.23	76.08	59.07	78.42	61.89	77.32	56.63	79.23	57.40	78.60				
MaxLogit [11]	59.11	79.21	51.83	82.90	52.94	78.91	53.90	81.65	62.39	78.39	57.67	79.75	56.31	80.14				
ODIN [20]	60.63	78.18	55.21	81.63	<b>45.94</b>	<b>83.79</b>	67.43	74.54	62.37	79.34	59.73	79.45	58.55	79.49				
IODIN [27]	59.09	79.24	51.57	82.96	52.93	78.89	54.06	81.56	62.07	78.48	57.47	79.83	56.20	80.16				
KL-Matching [12]	84.77	73.92	70.99	79.21	72.88	74.15	50.31	79.32	81.80	75.76	81.62	75.68	73.73	76.34				
Mahalanobis [19]	88.00	55.87	79.04	61.50	71.71	67.47	67.22	70.67	70.49	76.26	79.60	63.15	76.01	65.82				
ASH [7]	68.07	76.47	63.37	79.92	66.58	77.23	<b>45.98</b>	<b>85.60</b>	61.29	<b>80.72</b>	62.96	78.76	61.38	79.78				
SHE [39]	60.41	78.15	57.73	79.74	58.78	76.76	59.16	80.97	73.28	73.64	65.26	76.30	62.44	77.59				
LogitGap	<b>58.70</b>	<b>79.43</b>	<b>50.01</b>	<b>83.31</b>	52.49	78.60	54.84	81.07	<b>60.34</b>	78.86	<b>56.12</b>	<b>80.41</b>	<b>55.42</b>	<b>80.28</b>				
OE [15]	63.87	74.64	<b>0.41</b>	<b>99.88</b>	35.03	91.28	56.24	85.73	52.83	83.98	60.49	76.89	44.81	85.40				
+LogitGap	<b>63.29</b>	<b>75.73</b>	0.42	<b>99.88</b>	<b>24.96</b>	<b>92.93</b>	<b>50.28</b>	<b>87.68</b>	<b>49.82</b>	<b>85.37</b>	<b>58.67</b>	<b>78.21</b>	<b>41.24</b>	<b>86.63</b>				
MixOE [38]	<b>61.09</b>	78.18	49.43	83.93	68.04	<b>70.06</b>	76.72	73.06	66.37	78.03	56.10	80.44	62.96	77.28				
+LogitGap	61.31	<b>78.41</b>	<b>48.41</b>	<b>84.05</b>	<b>67.43</b>	69.98	<b>75.10</b>	<b>73.22</b>	<b>65.58</b>	<b>78.22</b>	<b>55.31</b>	<b>80.69</b>	<b>62.19</b>	<b>77.43</b>				

### 4.2 Results on Traditional OOD Detection with ImageNet as ID Dataset

In the traditional OOD detection setting, we further evaluate the performance of LogitGap using a ResNet-50 [10] backbone on the large-scale ImageNet [5] dataset. Following the OpenOOD [36] protocol, we adopt SSB-hard [33] and NINCO [2] as **near-OOD datasets**, iNaturalist[32], Textures[3], and OpenImage-O [35] as **far-OOD datasets**. In this setup, the ResNet-50 model [10] is trained from scratch on the ID data using cross-entropy loss. We compare LogitGap against several representative

Table 4: OOD Detection performance on ResNet-50 using ImageNet as the ID dataset.

Method	Near OOD				Far OOD						AVG	
	SSB-hard		NINCO		iNaturalist		Textures		OpenImage-O		FPR95	AUROC
	FPR95	AUROC	FPR95	AUROC	FPR95	AUROC	FPR95	AUROC	FPR95	AUROC		
MSP [14]	<b>73.6</b>	<b>73.2</b>	54.13	82.05	29.55	91.70	50.59	87.21	37.44	88.94	49.06	84.62
MaxLogit [11]	76.19	72.51	59.49	80.41	30.59	91.16	46.12	88.39	37.88	<b>89.17</b>	50.05	84.33
KL-Matching [12]	84.72	71.38	60.28	81.90	38.51	90.79	52.38	84.72	48.94	87.30	56.97	83.22
ODIN [20]	76.86	71.74	68.08	77.77	36.12	91.16	49.25	89.01	46.48	88.23	55.36	83.58
GradNorm [17]	78.24	71.90	79.50	74.02	32.01	<b>93.89</b>	<b>43.24</b>	<b>92.05</b>	68.46	84.83	60.29	83.34
Energy [21]	76.53	72.08	60.61	79.70	31.33	90.63	45.77	88.70	38.07	89.06	50.46	84.03
LogitGap	75.48	72.51	<b>53.64</b>	<b>82.25</b>	<b>26.68</b>	92.24	47.11	87.15	<b>35.10</b>	89.11	<b>47.60</b>	<b>84.65</b>

Table 5: OOD Detection performance on CLIP ViT-B/16 under zero-shot setting. Results are reported with ImageNet as the ID dataset in the far-OOD scenario.

Method	iNaturalist			OOD Dataset Textures			OpenImage-O			AVG		
	FPR95 ↓	AUROC ↑	AUPR ↑	FPR95 ↓	AUROC ↑	AUPR ↑	FPR95 ↓	AUROC ↑	AUPR ↑	FPR95 ↓	AUROC ↑	AUPR ↑
	Zero-shot											
Energy [21]	73.86	87.08	97.28	92.71	66.08	94.77	65.60	85.91	94.55	77.39	79.69	95.53
MCM [23]	31.49	94.39	98.80	58.76	<b>85.84</b>	<b>98.02</b>	40.76	91.99	96.95	43.67	90.74	97.92
MaxLogit [11]	56.25	90.47	97.99	86.45	71.85	95.66	51.89	88.93	95.66	64.86	83.75	96.44
LogitGap	32.54	94.13	98.74	58.56	85.73	97.96	39.12	92.41	97.11	43.41	90.76	97.94
LogitGap*	<b>27.82</b>	<b>94.89</b>	<b>98.91</b>	<b>58.17</b>	85.68	97.97	<b>37.27</b>	<b>92.66</b>	<b>97.19</b>	<b>41.09</b>	<b>91.08</b>	<b>98.02</b>

post-hoc OOD detection methods, including MSP [14], MaxLogit [11], KL-Matching [12], ODIN [20], IODIN [27], Energy [21], and GradNorm [17].

As summarized in Table 4, LogitGap achieves comparable or superior performance across all benchmarks and obtains the best overall average performance, demonstrating its robustness and effectiveness in the traditional OOD detection scenario.

#### 4.3 Results on Far-OOD Evaluation with ImageNet as ID Dataset in Zero-Shot OOD Detection

In the main paper, to isolate the effect of semantic shift, we construct a challenging OOD detection benchmark specifically designed for semantic shift evaluation. This benchmark is built using ImageNet [5] as ID dataset, while NINCO [2], ImageNet-O [16], and ImageNet-OOD [37] as OOD datasets, enabling a more accurate assessment of the model’s ability to detect semantic-level distribution changes.

Moreover, following the OpenOOD [36] evaluation protocol, we assess the performance of OOD detection methods in the far-OOD scenario under a zero-shot setting. In the far-OOD scenario, ImageNet is used as the ID dataset, while iNaturalist [32], OpenImage-O [35], and Textures [3] serve as the OOD datasets. As shown in Table 5, LogitGap maintains robust performance under the broader distributional shift.

#### 4.4 Results on More Architectures with ImageNet as ID Dataset in Zero-Shot OOD Detection

We extend our experiments to other CLIP variants with ImageNet as ID dataset, including ResNet-50, ResNet-101, ViT-B/32 and ViT-L/14 in zero-shot OOD detection. As shown in Table 6, LogitGap consistently outperforms baselines across all backbones. For instance, on ViT-L/14, LogitGap reduces FPR95 by 6.03% and improves AUROC by 2.65% compared to MCM. This demonstrate that LogitGap generalizes well across architectures.

Table 6: OOD Detection performance across various CLIP architectures under the zero-shot setting. Results are reported with ImageNet as the ID dataset in a semantic shift scenario.

Method	NINCO			OOD Dataset ImageNet-O			ImageNetOOD			AVG		
	FPR95 ↓	AUROC ↑	AUPR ↑	FPR95 ↓	AUROC ↑	AUPR ↑	FPR95 ↓	AUROC ↑	AUPR ↑	FPR95 ↓	AUROC ↑	AUPR ↑
Energy [21]	92.21	66.35	94.67	86.00	70.98	98.32	80.87	74.85	82.57	86.36	70.73	91.85
MCM [23]	<b>82.09</b>	71.58	95.55	84.25	73.95	98.57	84.97	75.19	83.64	83.77	73.57	92.59
MaxLogit [11]	88.19	69.42	95.23	83.95	72.50	98.41	<b>78.74</b>	<b>76.64</b>	83.91	83.63	72.85	92.52
LogitGap	82.65	<b>73.26</b>	<b>95.87</b>	<b>83.70</b>	<b>75.44</b>	<b>98.67</b>	84.01	76.51	<b>84.35</b>	<b>83.45</b>	<b>75.07</b>	<b>92.96</b>

(a) Results on ResNet-50.

Method	NINCO			OOD Dataset ImageNet-O			ImageNetOOD			AVG		
	FPR95 ↓	AUROC ↑	AUPR ↑	FPR95 ↓	AUROC ↑	AUPR ↑	FPR95 ↓	AUROC ↑	AUPR ↑	FPR95 ↓	AUROC ↑	AUPR ↑
Energy [21]	91.83	66.06	94.62	86.15	70.51	98.32	81.19	75.07	83.21	86.39	70.55	92.05
MCM [23]	84.13	70.66	95.38	83.75	74.06	98.56	85.19	74.76	83.11	84.36	73.16	92.35
MaxLogit [11]	87.94	69.41	95.24	84.15	72.46	98.43	79.73	<b>76.83</b>	<b>84.37</b>	83.94	72.90	92.68
LogitGap	<b>82.65</b>	<b>73.36</b>	<b>95.89</b>	<b>82.10</b>	<b>75.94</b>	<b>98.69</b>	<b>83.99</b>	76.18	83.88	<b>82.91</b>	<b>75.16</b>	<b>92.82</b>

(b) Results on ResNet-101.

Method	NINCO			OOD Dataset ImageNet-O			ImageNetOOD			AVG		
	FPR95 ↓	AUROC ↑	AUPR ↑	FPR95 ↓	AUROC ↑	AUPR ↑	FPR95 ↓	AUROC ↑	AUPR ↑	FPR95 ↓	AUROC ↑	AUPR ↑
Energy [21]	87.05	68.75	94.96	83.80	72.46	98.39	79.66	75.12	82.67	83.50	72.11	92.01
MCM [23]	79.62	73.85	95.93	80.95	75.58	98.66	84.09	76.32	84.16	81.55	75.25	92.92
MaxLogit [11]	82.27	71.76	95.53	80.55	74.32	98.51	<b>77.38</b>	76.99	84.06	<b>80.07</b>	74.36	92.70
LogitGap	<b>79.13</b>	<b>75.66</b>	<b>96.27</b>	<b>79.15</b>	<b>77.45</b>	<b>98.77</b>	82.03	<b>77.68</b>	<b>84.84</b>	80.10	<b>76.93</b>	<b>93.29</b>

(c) Results on ViT-B/32.

Method	NINCO			OOD Dataset ImageNet-O			ImageNetOOD			AVG		
	FPR95 ↓	AUROC ↑	AUPR ↑	FPR95 ↓	AUROC ↑	AUPR ↑	FPR95 ↓	AUROC ↑	AUPR ↑	FPR95 ↓	AUROC ↑	AUPR ↑
Energy [21]	81.88	76.89	96.51	77.45	78.45	98.81	76.95	78.16	84.61	78.76	77.83	93.31
MCM [23]	69.48	79.13	96.74	68.30	82.48	99.06	73.99	80.71	86.60	70.59	80.77	94.13
MaxLogit [11]	74.23	79.18	96.86	71.85	80.73	98.94	72.78	79.95	85.69	72.95	79.95	93.83
LogitGap	<b>64.43</b>	<b>82.49</b>	<b>97.38</b>	<b>62.05</b>	<b>84.94</b>	<b>99.20</b>	<b>67.20</b>	<b>82.84</b>	<b>87.75</b>	<b>64.56</b>	<b>83.42</b>	<b>94.78</b>

(d) Results on ViT-L/14.

## 5 More Analysis about LogitGap

### 5.1 Generalization Beyond Visual Tasks

Although LogitGap is primarily designed for visual classification, one of the core settings for OOD detection, we further verify its generalization to non-visual tasks. Specifically, we conduct an evaluation on the ESC-50 [25] audio classification dataset, where 50 categories are randomly divided into 25 ID and 25 OOD classes. The method is implemented with the CLAP [9] model, following the same hyperparameter selection strategy used in the visual domain, with  $N = 20\%$  of the class count ( $N = 5$ ).

As shown in Table 7, LogitGap maintains strong performance on this audio task, indicating that it does not rely on modality-specific architectures or features. This confirms its broad applicability across domains. In addition, we introduce MCM\_topN, a comparative baseline that applies the top-N strategy to MCM [23]. For fair comparison, we use the same  $N$  for MCM\_topN and LogitGap. Results show that applying the top-N strategy to MCM does not consistently improve performance, which indicates LogitGap’s advantage goes beyond top-N filtering.

Table 7: OOD Detection performance on CLAP using ESC-50 dataset, where 50 classes are randomly split into 25 ID and 25 OOD classes.

	FPR95 ↓	AUROC ↑	AUPR ↑
MCM [23]	26.40	94.86	95.35
MaxLogit [11]	40.60	92.32	92.51
Energy [21]	42.10	91.76	92.09
MCM_topN	27.90	94.69	95.21
GEN [22]	24.10	95.55	95.96
LogitGap	<b>17.50</b>	<b>96.15</b>	<b>96.48</b>

### 5.2 Relationship Between LogitGap and Other Logit-Pattern-Based Methods

We review related works in both active learning and OOD detection, and summarize several representative methods in Table 8. Here,  $z_c$  denotes the logit and  $p_c$  represents the predicted probability for  $c$ -th class. Specifically, we compare LogitGap with three classical active learning approaches: LC [4], Margin [29], and Entropy [30]; and four representative OOD detection methods: MSP [14], MaxLogit [11], DML [40], and GEN [22].

Among these methods, Margin Sampling is the most closely related to ours, as it also measures a margin between model outputs. However, two key distinctions set LogitGap apart: (1) Representation Level: Margin Sampling operates on predicted probabilities, whereas LogitGap computes directly on raw logits, avoiding softmax-induced normalization effects. (2) Scope of Comparison: Margin Sampling considers only the top-2 predictions, while LogitGap generalizes this to the top-N logits, enabling a more holistic uncertainty estimation.

Table 8: Comparison of logit-pattern-based methods, “OOD” and “AL” represent OOD detection and active learning respectively.

Method	Task	Equation
MSP [14]	OOD	$\max_c p_c$
MaxLogit [11]	OOD	$\max_c z_c$
DML [40]	OOD	$\max_c \lambda \hat{z}_c +   z_c  , z_c = \hat{z}_c \cdot   z_c  $
GEN [22]	OOD	$-\sum_{c=1}^M p_c^\gamma (1 - p_c)^\gamma$ , and $p_1 \geq \dots \geq p_M \geq \dots \geq p_N, \gamma \in (0, 1)$
Margin [29]	AL	$p_1 - p_2$ and $p_1 \geq p_2 \dots \geq p_N$
LC [4]	AL	$\max_c 1 - p_c$
Entropy [30]	AL	$-\sum_c p_c \cdot \log p_c$
LogitGap	OOD	$\frac{1}{M-1} \sum_{c=2}^M z_1 - z_c$ and $z_1 \geq \dots \geq z_M \geq \dots \geq z_N$

Table 9: OOD Detection Performance on CLIP ViT-B/16 under Zero-shot Setting. Results are reported with ImageNet as the ID dataset in a semantic shift scenario.

Method	NINCO			OOD Dataset ImageNet-O			ImageNetOOD			AVG		
	FPR95 ↓	AUROC ↑	AUPR ↑	FPR95 ↓	AUROC ↑	AUPR ↑	FPR95 ↓	AUROC ↑	AUPR ↑	FPR95 ↓	AUROC ↑	AUPR ↑
Margin Sampling	89.88	68.40	94.97	89.45	67.83	98.14	89.96	67.49	77.67	89.76	67.91	90.26
LogitGap	<b>77.42</b>	<b>76.51</b>	<b>96.38</b>	<b>71.95</b>	<b>81.45</b>	<b>99.03</b>	<b>75.40</b>	<b>80.27</b>	<b>86.21</b>	<b>74.92</b>	<b>79.41</b>	<b>93.87</b>

Table 10:  $N$  selected using synthetic and real OOD samples on ViT-B/16 with ImageNet as ID dataset.  $N_{syn}$  and  $N_{real}$  denote  $N$  selected using synthetic OOD and real OOD samples, respectively.

	NINCO			ImageNet-O			ImageNetOOD		
	FPR95 ↓	AUROC ↑	$N$	FPR95 ↓	AUROC ↑	$N$	FPR95 ↓	AUROC ↑	$N$
$N_{syn}$	77.42	76.51	88	71.95	81.45	88	75.40	80.27	88
$N_{real}$	77.22	76.51	100	71.60	81.50	110	75.39	80.27	90

To emphasize the importance of these distinctions, we adapt Margin Sampling for OOD detection under zero-shot setting using CLIP ViT-B/16, with ImageNet as the ID dataset in a semantic shift scenario. As shown in Table 9, LogitGap consistently outperforms Margin Sampling across all benchmarks, demonstrating the effectiveness and robustness of our formulation.

### 5.3 The Effectiveness of Synthetic OOD Data

As described in Section 3.3, we propose an OOD data synthesis strategy to adaptively select the hyperparameter  $N$ . Since the synthesized OOD data are derived from in-distribution (ID) information, they may not fully capture the characteristics of real-world OOD data. Nevertheless, our empirical findings indicate that such synthetic samples are sufficiently informative for selecting a robust hyperparameter  $N$ . As shown in Table 10, the optimal  $N$  determined using synthetic OOD samples (*i.e.*,  $N_{syn}$ ) is highly consistent with the one obtained from multiple real OOD datasets (*i.e.*,  $N_{real}$ ), achieving comparable detection performance. These results validate the practicality of the  $N$ -selection strategy, even in the absence of real OOD data.

### 5.4 The Impact of Hyperparameter $N$ on LogitGap Performance

In Table 11, we provide a more ablation study on hyperparameter  $N$ . Specifically, we conduct experiments under the zero-shot OOD detection setting, using either ImageNet-100 or ImageNet-1K as ID dataset. For simplicity, we report the FPR95 of our LogitGap method based on CLIP ViT-B/16 model. We can observe that: (1) Optimal  $N$  varies by dataset (e.g., 19 for ImageNet-100, 195 for ImageNet-1K); (2) Setting  $N$  to 20% of total classes consistently provides strong performance. Therefore, we adopt this value as default in LogitGap.

Table 11: Effect of hyperparameter  $N$  on FPR95 using ViT-B/16 under zero-shot setting.

	5	95	195	295	395	495	595	695	795	895	995
NINCO	83.26	77.34	76.81	76.56	76.40	76.74	77.00	77.53	78.29	78.68	79.65
ImageNet-O	81.80	71.85	72.45	73.20	73.40	73.55	73.95	74.30	74.85	75.50	75.90
ImageNetOOD	82.49	75.47	76.38	77.16	77.64	78.26	78.74	79.19	79.72	80.31	81.04

(a) ImageNet as ID dataset.

	1	9	19	29	39	49	59	69	79	89	99
NINCO	75.54	46.27	42.77	41.88	41.85	42.6	43.23	44.49	46.31	47.64	50.58
ImageNet-O	76.00	45.70	43.50	44.25	44.30	45.25	45.75	46.45	47.90	47.90	48.65
ImageNetOOD	76.02	46.86	46.19	46.62	47.10	47.80	48.33	49.32	50.51	50.57	51.53

(b) ImageNet-100 as ID dataset.

## References

- [1] Yichen Bai, Zongbo Han, Changqing Zhang, Bing Cao, Xiaoheng Jiang, and Qinghua Hu. Id-like prompt learning for few-shot out-of-distribution detection. In *CVPR*, 2024.
- [2] Julian Bitterwolf, Maximilian Müller, and Matthias Hein. In or out? fixing imagenet out-of-distribution detection evaluation. In *ICML*, 2023.
- [3] Mircea Cimpoi, Subhransu Maji, Iasonas Kokkinos, Sammy Mohamed, and Andrea Vedaldi. Describing textures in the wild. In *CVPR*, 2014.
- [4] Aron Culotta and Andrew McCallum. Reducing labeling effort for structured prediction tasks. In *AAAI*, 2005.
- [5] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *CVPR*, 2009.
- [6] Li Deng. The mnist database of handwritten digit images for machine learning research [best of the web]. *IEEE signal processing magazine*, 29(6):141–142, 2012.
- [7] Andrija Djuricic, Nebojsa Bozanic, Arjun Ashok, and Rosanne Liu. Extremely simple activation shaping for out-of-distribution detection. In *ICLR*, 2023.
- [8] 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. In *ICLR*, 2021.
- [9] Benjamin Elizalde, Soham Deshmukh, Mahmoud Al Ismail, and Huaming Wang. Clap learning audio concepts from natural language supervision. In *ICASSP*, 2023.
- [10] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, 2016.
- [11] Dan Hendrycks, Steven Basart, Mantas Mazeika, Mohammadreza Mostajabi, Jacob Steinhardt, and Dawn Song. Scaling out-of-distribution detection for real-world settings. In *ICML*, 2022.
- [12] Dan Hendrycks, Steven Basart, Mantas Mazeika, Andy Zou, Joe Kwon, Mohammadreza Mostajabi, Jacob Steinhardt, and Dawn Song. Scaling out-of-distribution detection for real-world settings. In *ICML*, 2022.
- [13] Dan Hendrycks, Steven Basart, Norman Mu, Saurav Kadavath, Frank Wang, Evan Dorundo, Rahul Desai, Tyler Zhu, Samyak Parajuli, Mike Guo, et al. The many faces of robustness: A critical analysis of out-of-distribution generalization. In *CVPR*, 2021.
- [14] Dan Hendrycks and Kevin Gimpel. A baseline for detecting misclassified and out-of-distribution examples in neural networks. In *ICLR*, 2017.
- [15] Dan Hendrycks, Mantas Mazeika, and Thomas Dietterich. Deep anomaly detection with outlier exposure. In *ICLR*, 2019.
- [16] Dan Hendrycks, Kevin Zhao, Steven Basart, Jacob Steinhardt, and Dawn Song. Natural adversarial examples. In *CVPR*, 2021.
- [17] Rui Huang, Andrew Geng, and Yixuan Li. On the importance of gradients for detecting distributional shifts in the wild. In *NeurIPS*, 2021.
- [18] A. Krizhevsky and G. Hinton. Learning multiple layers of features from tiny images. *Handbook of Systemic Autoimmune Diseases*, 1(4), 2009.
- [19] Kimin Lee, Kibok Lee, Honglak Lee, and Jinwoo Shin. A simple unified framework for detecting out-of-distribution samples and adversarial attacks. In *NeurIPS*, 2018.
- [20] Shiyu Liang, Yixuan Li, and Rayadurgam Srikant. Enhancing the reliability of out-of-distribution image detection in neural networks. In *ICLR*, 2018.

- [21] Weitang Liu, Xiaoyun Wang, John D Owens, and Yixuan Li. Energy-based out-of-distribution detection. In *NeurIPS*, 2020.
- [22] Xixi Liu, Yaroslava Lochman, and Zach Christopher. Gen: Pushing the limits of softmax-based out-of-distribution detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023.
- [23] Yifei Ming, Ziyang Cai, Jiuxiang Gu, Yiyu Sun, Wei Li, and Yixuan Li. Delving into out-of-distribution detection with vision-language representations. In *NeurIPS*, 2022.
- [24] Yuval Netzer, Tao Wang, Adam Coates, Alessandro Bissacco, Baolin Wu, Andrew Y Ng, et al. Reading digits in natural images with unsupervised feature learning. In *NeurIPS Workshop*, 2011.
- [25] Karol J. Piczak. ESC: Dataset for Environmental Sound Classification. In *ACM Multimedia*, 2015.
- [26] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *ICML*, 2021.
- [27] Sudarshan Regmi. Going beyond conventional ood detection. *arXiv preprint arXiv:2411.10794*, 2024.
- [28] Tal Ridnik, Emanuel Ben-Baruch, Asaf Noy, and Lihi Zelnik-Manor. Imagenet-21k pretraining for the masses. In *NeurIPS*, 2021.
- [29] Dan Roth and Kevin Small. Margin-based active learning for structured output spaces. In *Proceedings of the 17th European Conference on Machine Learning*, 2006.
- [30] Burr Settles and Mark Craven. An analysis of active learning strategies for sequence labeling tasks. In *EMNLP*, 2008.
- [31] Antonio Torralba, Rob Fergus, and William T Freeman. 80 million tiny images: A large data set for nonparametric object and scene recognition. *TPAMI*, 2008.
- [32] Grant Van Horn, Oisin Mac Aodha, Yang Song, Yin Cui, Chen Sun, Alex Shepard, Hartwig Adam, Pietro Perona, and Serge Belongie. The inaturalist species classification and detection dataset. In *CVPR*, 2018.
- [33] Sagar Vaze, Kai Han, Andrea Vedaldi, and Andrew Zisserman. Open-set recognition: a good closed-set classifier is all you need? In *ICLR*, 2022.
- [34] Haohan Wang, Songwei Ge, Zachary Lipton, and Eric P Xing. Learning robust global representations by penalizing local predictive power. In *NeurIPS*, 2019.
- [35] Haoqi Wang, Zhizhong Li, Litong Feng, and Wayne Zhang. Vim: Out-of-distribution with virtual-logit matching. In *CVPR*, 2022.
- [36] Jingkan Yang, Pengyun Wang, Dejian Zou, Zitang Zhou, Kunyuan Ding, Wenxuan Peng, Haoqi Wang, Guangyao Chen, Bo Li, Yiyu Sun, et al. Openood: Benchmarking generalized out-of-distribution detection. In *NeurIPS*, 2022.
- [37] William Yang, Byron Zhang, and Olga Russakovsky. Imagenet-ood: Deciphering modern out-of-distribution detection algorithms. In *ICLR*, 2024.
- [38] Jingyang Zhang, Nathan Inkawhich, Randolph Linderman, Yiran Chen, and Hai Li. Mixture outlier exposure: Towards out-of-distribution detection in fine-grained environments. In *WACV*, 2023.
- [39] Jinsong Zhang, Qiang Fu, Xu Chen, Lun Du, Zelin Li, Gang Wang, Shi Han, Dongmei Zhang, et al. Out-of-distribution detection based on in-distribution data patterns memorization with modern hopfield energy. In *ICLR*, 2023.

- [40] Zihan Zhang and Xiang Xiang. Decoupling maxlogit for out-of-distribution detection. In *CVPR*, 2023.
- [41] 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*, 2017.