# Mining In-distribution Attributes in Outliers for Out-of-distribution Detection

**Yutian Lei, Luping Ji*, Pei Liu**

School of Computer Science and Engineering, University of Electronic Science and Technology of China, Chengdu, China
ytlei823@gmail.com, jiluping@uestc.edu.cn, yuukilp@std.uestc.edu.cn

## Abstract

Out-of-distribution (OOD) detection is indispensable for deploying reliable machine learning systems in real-world scenarios. Recent works, using auxiliary outliers in training, have shown good potential. However, they seldom concern the intrinsic correlations between in-distribution (ID) and OOD data. In this work, we discover an obvious correlation that OOD data usually possesses significant ID attributes. These attributes should be factored into the training process, rather than blindly suppressed as in previous approaches. Based on this insight, we propose a structured multi-view-based out-of-distribution detection learning (MVOL) framework, which facilitates rational handling of the intrinsic in-distribution attributes in outliers. We provide theoretical insights on the effectiveness of MVOL for OOD detection. Extensive experiments demonstrate the superiority of our framework to others. MVOL effectively utilizes both auxiliary OOD datasets and even wild datasets with noisy ID data.

**Code** — https://github.com/UESTC-nnLab/MVOL

## Introduction

Modern deep neural networks could produce overconfident and unreliable predictions when their test inputs are out-of-distribution (OOD) (Nguyen, Yosinski, and Clune 2015). This presents a crucial challenge, especially for deploying deep learning models in the real world. To tackle this challenge, recent studies have explored outlier exposure (OE)-based training strategies (Hendrycks, Mazeika, and Dietterich 2019). They utilize an auxiliary outlier dataset in training to suppress the model's response to outliers, *i.e.*, out-of-distribution data, thus detecting those inputs far away from in-distribution (ID) data (Chen et al. 2021; Ming, Fan, and Li 2022; Zhu et al. 2023). These strategies have achieved considerable success and often perform better than those without auxiliary data (Du et al. 2022; Djurisic et al. 2023).

Despite the promising results, existing OE-based methods pay limited attention to what underlying correlations exist between ID and OOD data, still suffering from the irrational use of auxiliary outliers. As shown in Figure 1(a), for models trained solely on ID data, we find they tend to present higher logits (*i.e.*, pre-softmax output) on specific known categories than others when their inputs are out-of-distribution. This finding implies that outliers could contain ID attributes, although they are often believed to be semantically distinct from in-distribution (Bai et al. 2023). However, previous approaches, *e.g.*, outlier exposure (Hendrycks, Mazeika, and Dietterich 2019), and energy-regularized learning (Liu et al. 2020), usually ignore those intrinsic ID attributes. Specifically, they treat these attributes as entirely random noise and attempt to suppress models' responses to them blindly. Such behavior indicates unreasonable use of auxiliary outliers and may degrade the detection performance. The above discrepancy naturally motivates the following question: *how can we rationally handle the intrinsic ID attributes in auxiliary outlier data for OOD detection?*

To address this question, we propose a structured multi-view-based out-of-distribution learning framework (**MVOL**) in Figure 1 (d) — tackling OOD detection via mining ID attributes in outliers. *In data level*, this framework involves an extended multi-view data model. It establishes ID and OOD data in a unified perspective, revealing intrinsic in-distribution attributes in outliers. *In model level*, MVOL employs maximum logit (MaxLogit) as the OOD score with new insight, which naturally adapts to our data model. To calibrate unexpected high MaxLogit produced for outlier input, MVOL involves a multi-view-based learning objective to rationally utilize ID attributes in auxiliary outliers.

**Extended Multi-view Data Model (MVDM).** The original MVDM (Allen-Zhu and Li 2023) posits that in a natural dataset, a data point comprises main features, minor features, and noise [1]. To formally study the correlations between ID and OOD data, we extend the original MVDM to the context of out-of-distribution detection by assuming that *outliers mainly consist of minor ID features and noise*. This new assumption originates from our two empirical findings. **(1)** Outliers contain ID attributes, as indicated by Figure 1(a) and analyzed above. **(2)** These ID attributes generally have lower magnitudes than those in ID data. As shown in Figure 1 (b), by comparing averages of maximum logit on ID

*Corresponding author.



[1] In a vision dataset featuring car and cat categories, a car image typically comprises headlights, wheels, or windows, termed main features. Additionally, some cars might have headlights resembling cat eyes, which can also be recognized as a small fraction of "cat features", termed minor features.

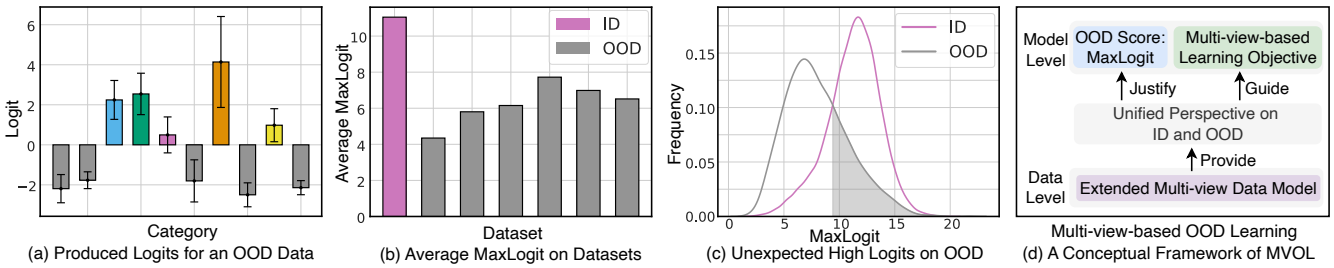



Figure 1: (a)-(c) Motivation and (d) overview of MVOL. (a) OOD data could have ID attributes. The Y-axis represents average logits on ten models trained independently, (b) outliers mainly have minor ID features, and (c) the model trained solely on ID data can produce over-activated logits for certain outliers. The gray-shaded area highlights those outliers with high probability.

and six OOD datasets, we observe that ID data has a significantly larger value. This observation suggests that ID attributes in outliers correspond to minor ID features. This is because compared to main ID features, minor ones typically carry a smaller weight, leading a well-trained model to produce smaller logit responses, as shown in MVDM.

**New Insight on MaxLogit as an OOD Score.** Beyond empirical evidence in previous works (Hendrycks et al. 2022; Vaze et al. 2022), our extended MVDM justifies that MaxLogit serves as an interpretable and effective OOD score. (1) **Interpretable**: MaxLogit measures ID attributes contained in test inputs. Outliers with minor ID features tend to have a low score. (2) **Effective**: When a well-calibrated model learns all ID features, its error bound is near zero theoretically. Nevertheless, for a model trained solely on ID data, it could still be uncalibrated and produce unexpected high MaxLogit for outlier inputs, as shown in Figure 1(c), which degrades the detection performance (Wei et al. 2022).

**Multi-view-based Learning Objective.** We propose a learning objective for OOD detection — explicitly utilizing the intrinsic minor ID features in auxiliary outliers to calibrate unexpected high logits. This is in contrast with confidence loss (Lee et al. 2018a) in OE, which forces models to respond *uniformly* to all ID categories for outliers. It causes models to underemphasize minor ID features yet overemphasize noise, thus violating the heuristics that minor ID features deserve larger logit responses than noise. Accordingly, our learning process encourages models to more accurately respond to the minor ID features and noise in outliers.

**Contributions: (1)** To our knowledge, we are among the first works that explicitly mine in-distribution attributes in outliers to tackle out-of-distribution detection. **(2)** We propose a multi-view-based out-of-distribution learning framework (**MVOL**) to handle the intrinsic ID attributes in outliers rationally. **(3)** We show that MVOL performs overall strong performance while training with OOD datasets and even wild datasets with ID data as noise.

## Related Work

**Out-of-distribution Detection.** (Hendrycks and Gimpel 2017) presents a baseline for OOD detection using a pre-trained model's maximum softmax probability. Subsequent works develop different scoring functions, such as ODIN (Liang, Li, and Srikant 2018), Mahalanobis (Lee et al. 2018b), Energy (Liu et al. 2020), MaxLogit (Hendrycks et al. 2022; Vaze et al. 2022). Some methods synthesize outliers to regularize models. Synthesized outliers are represented in feature space (Du et al. 2022; Tao et al. 2023) or pixel space (Zheng et al. 2023; Du et al. 2023).

Another promising line of work uses auxiliary outliers to regularize models. Outlier exposure (Hendrycks, Mazeika, and Dietterich 2019) using confidence loss (Lee et al. 2018a) encourages models to output a uniform distribution of softmax probability on outliers. Energy-based regularization (Energy) (Liu et al. 2020) trains models to widen the energy gap by increasing the sum of logits for ID samples above a lower bound and reducing those of outliers below an upper bound. Both OE and Energy fail to consider ID and OOD correlations explicitly. Some other works highlight using informative outliers to regularize models (Chen et al. 2021; Ming, Fan, and Li 2022; Zhu et al. 2023). They often employ OE and Energy-based learning objectives and fail to utilize correlations between ID and OOD data in regularization. In addition, (Yang et al. 2021; Katz-Samuels et al. 2022) explore a more complex setting where auxiliary datasets could contain ID noise. Strong ID attributes exist here. Our proposed method also shows good robustness in this case.

**Multi-view Data Model (MVDM).** (Allen-Zhu and Li 2023) presents the MVDM to explore ensemble learning. It assumes a natural data point has multiple views that can be used for classification. A single neural network trained with gradient descent might classify a data point based on only one of those views. Ensembling multiple models or distilling ensemble models into one model could uncover all features. Previous research in OOD detection with auxiliary outliers usually characterizes ID and OOD data in latent space using a simplified Gaussian-like model (Chen et al. 2021). This model makes it hard to understand the correlations between ID and OOD data concretely. Instead, we extend MVDM to provide a systematic understanding of their correlations.

## Preliminaries on Multi-view Data Model

We first give the formal in-distribution definition by restating the original Multi-view Data Model (Allen-Zhu and Li 2023) in OOD detection context. Moreover, since MVDM contains two typical settings, *i.e.*, single model and ensemble distillation, we will review them and briefly summarize how models perform feature learning in these two settings.

## In-distribution Definition and Notations

We consider a $k$-class classification setting in OOD detection. Any data input is represented by $P$ patches $X = (x_1, x_2, \ldots, x_P) \in \mathbb{R}^{d\times P}$, where each patch has dimension $d$. For ID data, labels belong to $[k]$. It is assumed that there are multiple associated features for each label $j \in [k]$, say two features for simplicity, represented by two *orthogonal unit feature vectors* $v_{j,1}, v_{j,2} \in \mathbb{R}^d$.

Our in-distribution definition is adapted from the original MVDM for OOD detection. $D^{in}$ is defined via the multi-view distribution $D_m^{in}$ and single-view distribution $D_s^{in}$. $D_m^{in}$ represents images that we can observe all main ID features and use any of these features to classify them. $D_s^{in}$ represents images taken from a particular angle, where one or more main ID features can be missing, *i.e.*,

**Definition 1** (data distributions $D_m^{in}$ and $D_s^{in}$)**.** *Given $D \in \{D_m^{in}, D_s^{in}\}$, we define $(X^{in}, y) \sim D$ as follows. First, choose the label $y \in [k]$ uniformly at random. Then, $X^{in}$ is generated as follows:*

***1.*** *Denote $\mathcal{V}(X^{in}) = \{v_{y,1}, v_{y,2}\} \cup \mathcal{V}'$ as the set of feature vectors used in this data vector $X$, where $\{v_{y,1}, v_{y,2}\}$ are **main ID features** and $\mathcal{V}'$ is a set of **minor ID features** uniformly sampled from $\{v_{j,1}, v_{j,2}\}_{j\in[k]\setminus\{y\}}$.*

***2.*** *For each $v \in \mathcal{V}(X)$, pick many disjoint patches in $[P]$ and denote them as $\mathcal{P}_v(X^{in}) \subset [P]$. We denote $\mathcal{P}(X^{in}) = \bigcup_{v\in\mathcal{V}(X^{in})} \mathcal{P}_v(X^{in})$.*

***3.*** *If $D = D_s^{in}$ is the single-view distribution, pick a value $\hat{\ell} = \ell(X^{in}) \in [2]$ uniformly at random.*

***4.*** *For each $v \in \mathcal{V}(X^{in})$ and $p \in \mathcal{P}_v(X^{in})$, $x_p = z_p v +$ "noise" $\in \mathbb{R}^d$. These random coefficients $z_p \geq 0$ satisfy:*

*· In the case of multi-view distribution $D = D_m^{in}$:*

*1) $\sum_{p\in\mathcal{P}_v(X^{in})} z_p \in [1, O(1)]$ when $v \in \{v_{y,1}, v_{y,2}\}$;*

*2) $\sum_{p\in\mathcal{P}_v(X^{in})} z_p \in [\Omega(1), 0.4]$ when $v \in V(X^{in}) \setminus \{v_{y,1}, v_{y,2}\}$.*

*· In the case of single-view distribution $D = D_s^{in}$:*

*1) $\sum_{p\in\mathcal{P}_v(X^{in})} z_p \in [1, O(1)]$ when picked $v = v_{y,\hat{\ell}}$;*

*2) $\sum_{p\in\mathcal{P}_v(X^{in})} z_p$ is much smaller than that of $v_{y,\hat{\ell}}$ and can be ignored when $v \in V(X^{in}) \setminus \{v_{y,\hat{\ell}}\}$.*

***5.*** *For each $p \in P \setminus P(X^{in})$, $x_p$ consists only of "noise".*

**Definition 2** ($D^{in}$ and $Z^{in}$)**.** *The final in-distribution $D^{in}$ consists of data from $D_m^{in}$ w.p. $\mu$ and $D_s^{in}$ w.p. $1-\mu$. We are given $N$ training samples from $D_{in}$, and denote the training data set as $Z^{in} = Z_m^{in} \cup Z_s^{in}$ where $Z_m^{in}$ and $Z_s^{in}$ respectively represent multi-view and single-view training data. We write $(X^{in}, y) \sim Z^{in}$ as $(X^{in}, y)$ sampled randomly from $Z^{in}$.*

Furthermore, a simplified neural network is provided to conduct analyses on MVDM. Concretely, this neural network is represented by a tow-layer convolutional network $F(X) = (F_1(X), \ldots, F_k(X)) \in \mathbb{R}^k$ parameterized by $w_{i,r} \in \mathbb{R}^d$. For $i \in [k], r \in [m]$, $w_{i,r}$ satisfies:

$$\forall i \in [k] : F_i(X) = \sum_{r\in[m]} \sum_{p\in[P]} \widetilde{\mathrm{ReLU}}(\langle w_{i,r}, x_p\rangle),$$

where $\widetilde{\mathrm{ReLU}}$ is a smoothed activation function.

## Feature-Learning Based on MVDM

To briefly introduce the learning mechanism under single model and ensemble distillation model settings, a thought experiment is utilized in (Allen-Zhu and Li 2023). Here, we present its basic settings and main conclusions to facilitate the understanding of our work.

**(1) Basic Settings:** Given $k = 2$ and four "features" $v_1, v_2, v_3, v_4$. $v \in \{v_1, v_2\}$ is associated with label 1, and $v' \in \{v_3, v_4\}$ is associated with 2. Main and minor ID features weigh 1 and 0.1, respectively. For each category, there are $80\%$ training data from $D_m^{in}$. The remaining $20\%$ are from $D_s^{in}$, *i.e.*, one half has one main ID feature, and the other half has the second. Data points with labels 1 and 2 could have minor ID features $v'$ and $v$, respectively.

**(2) Main Conclusions: (i) Single model** can learn only one feature of each category. While training a single neural network with random initialization using cross-entropy loss, the network will pick up one feature for each label and correctly classify $90\%$ training examples, *i.e.*, $80\%$ multi-view data and $10\%$ single-view data with the picked feature. The remaining $10\%$ examples will be memorized using noise in the data. Thus, this single model learns to classify each category using only one feature. **(ii) Ensemble model** can learn both features of each category. Depending on the random initialization, each network will pick up $v_1$ or $v_2$, each with a probability of $50\%$. Therefore, as long as we ensemble many independently trained models, with a high probability their ensemble will pick up all features $v_1, v_2, v_3, v_4$. **(iii) Ensemble distillation model** can learn both features of each category. Ensemble distillation (Hinton, Vinyals, and Dean 2015) trains an individual model to match the ensemble models' outputs. The key idea is that the model can learn all features via the soft labels produced by ensemble models.

# Method

In this section, we propose a structured multi-view-based out-of-distribution learning framework (**MVOL**) to tackle OOD detection via mining ID attributes in outliers.

## Extended Multi-view Data Model

We extend MVDM to study the correlations between ID and OOD data formally. OOD data is assumed to mainly consist of minor ID features and noise, *i.e.*,

**Definition 3** (Out-of-distribution $D^{out}$)**.** *We define $X^{out} \sim D^{out}$ as follows. $X^{out}$ is generated by:*

***1.*** *Denote $\mathcal{V}(X^{out})$ as the set of **minor ID feature** vectors used in this data vector $X^{out}$, which are uniformly sampled from $\{v_{j,1}, v_{j,2}\}_{j\in[k]}$.*

***2.*** *For each $v \in \mathcal{V}(X^{out})$, pick many disjoint patches in $[P]$ and denote it as $P_v(X^{out}) \subset [P]$. We denote $P(X^{out}) = \bigcup_{v\in\mathcal{V}(X^{out})} \mathcal{P}_v(X)$.*

***3.*** *For each $v \in \mathcal{V}(X^{out})$ and $p \in \mathcal{P}_v(X^{out})$, we set $x_p = z_p v +$ "noise" $\in \mathbb{R}^d$. These random coefficients $z_p \geq 0$ satisfy that: $\sum_{p\in\mathcal{P}_v(X^{out})} z_p \in [\Omega(1), 0.4]$.*

***4.*** *For each $p \in [P] \setminus \mathcal{P}(X^{out})$, $x_p$ consists only of "noise".*

**Definition 4** ($Z^{out}$)**.** *We are given $M$ auxiliary OOD training samples from $D^{out}$, and denote the training data set as*

*$Z^{out}$. We write $X^{out} \sim Z^{out}$ as $X^{out}$ sampled randomly from $Z^{out}$. We are given samples $M \geq N$ to represent a large auxiliary OOD dataset.*

## New Insight on MaxLogit as an OOD Score

MVOL employs MaxLogit as the OOD scoring function. MaxLogit measures ID attributes contained in test inputs. Outliers with minor ID features tend to have a low score. We justify MaxLogit with our extended MVDM.

**MaxLogit based OOD detector.** For a sample $(X^{in}, y) \sim D^{in}$ or $X^{out} \sim D^{out}$ and a neural network $F$. Feeding $X \in \{X^{in}, X^{out}\}$ into $F$, we get logit outputs $F(X) = (F_1(X), \ldots, F_k(X)) \in \mathbb{R}^k$. Then, the MaxLogit scoring function is given as follows.

$$\text{MaxLogit}(X; F) = \max(F_1(X), \ldots, F_k(X)).$$

Then MaxLogit can be used in the following OOD detector:

$$G(X; \tau, F) = \begin{cases} 0 & \text{if MaxLogit}(X; F) \leq \tau, \\ 1 & \text{if MaxLogit}(X; F) > \tau, \end{cases} \quad (1)$$

where 0 and 1 represent OOD and ID by convention, respectively. The threshold $\tau$ is chosen so that a high fraction of ID data (*e.g.*, 95%) is with the MaxLogit above $\tau$.

**Theoretical analysis with our extended MVDM.**

**(1) Interpretability of MaxLogit:** We demonstrate that MaxLogit can be an interpretable ID/OOD indicator, supported by our extended MVDM stated above. We define:

$$I(X) = \text{argmax}_{j \in [k]} \sum_{p \in P_{v_{j,1}}(X) \bigcup P_{v_{j,2}}(X)} z_p; \quad (2)$$

$$z(X) = \max_{j \in [k]} \sum_{p \in P_{v_{j,1}}(X) \bigcup P_{v_{j,2}}(X)} z_p \quad (3)$$

$I(X)$ is the category with the largest sum of coefficients on associated features. $z(X)$ is this sum value.

**Proposition 1.** *For every $X^{out} \sim D^{out}$, every $(X_s^{in}, y_s) \sim D_s^{in}$, and every $(X_m^{in}, y_m) \sim D_m^{in}$, we have:*

$$z(X^{out}) < z(X_s^{in}) \quad \textit{and} \quad z(X^{out}) < z(X_m^{in})$$

**Assumption 1.** *For every $i, j \in [k]$ and every $l, l^{'} \in [2]$, consider an ideal classifier $F$ satisfying*

$$\sum_{r \in [m]} [\langle w_{i,r}, v_{i,l} \rangle]^+ = \sum_{r \in [m]} [\langle w_{j,r}, v_{j,l^{'}} \rangle]^+ \pm o(1)$$

Note that throughout this paper $[a]^+ = \max\{0, a\}$.

**Intuition:** When Assumption 1 is valid, we can deduce that $F_{I(X^{out})}(X^{out}) < F_{I(X_s^{in})}(X_s^{in})$ and $F_{I(X^{out})}(X^{out}) < F_{I(X_m^{in})}(X_m^{in})$ with Proposition 1, which corresponds to relation among MaxLogit scores. In other words, the MaxLogit of ID data can be higher than that of OOD data. Therefore, MaxLogit serves as an interpretable OOD scoring function.

**(2) Efficacy of MaxLogit:** We provide error-bound analysis in the single model setting, where models learn only part of features for classification, and in the ensemble distillation model setting, where models uncover all features via distilling knowledge from ensemble models. These analyses reveal the efficacy of MaxLogit.

**Assumption 2.** *For every $i, j \in [k]$, every $l, l^{'} \in [2]$ and $v_{i,l}, v_{j,l'}$ are learned by single model, consider an ideal classifier $F$ satisfing:*

$$\sum_{r \in [m]} [\langle w_{i,r}, v_{i,l} \rangle]^+ = \sum_{r \in [m]} [\langle w_{j,r}, v_{j,l^{'}} \rangle]^+ \pm o(1)$$

**Theorem 1** (Calibrated Single Model)**.** *Suppose we train a single model from random initialization. When Assumption 2 is valid, we have:*

$$FNR(F) \leq \frac{1}{2}(1 - \mu + o(1))$$

**Theorem 2** (Calibrated Ensemble Distillation Model)**.** *Suppose we train a model using ensemble distillation. When Assumption 1 is valid, we have:*

$$FNR(F) \leq o(1)$$

Theorem 1 tells us **i)** the error bound in the single model setting can be established with the multi-view data proportion $\mu$ in ID datasets and **ii)** this bound decreases as $\mu$ increases. Theorem 2 shows that the error bound can reach near zero in the ensemble distillation model setting, where models learn all features. It highlights the superiority of MaxLogit as an OOD score.

## Multi-view-based Learning Objective

Models trained solely on in-distribution data could be uncalibrated and produce unexpected high MaxLogit for out-of-distribution inputs (Wei et al. 2022). We calibrate models' logits for OOD inputs with auxiliary outliers.

In this section, we first use confidence loss in OE as an example to analyze the common problem in current outlier exposure-based methods, *i.e.*, overlooking intrinsic ID attributes in outliers. Then a multi-view-based learning objective is proposed to utilize these attributes explicitly.

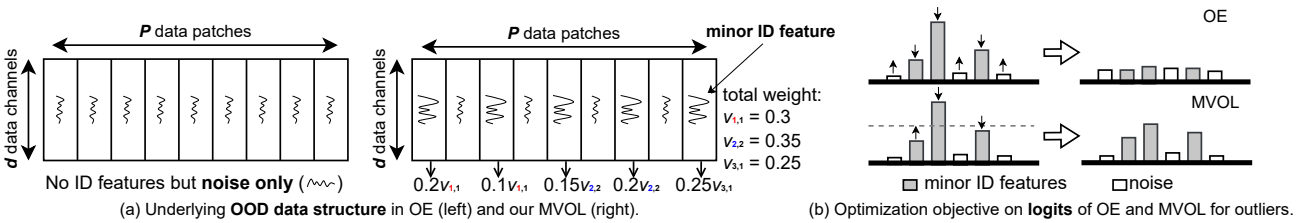
**Revisiting OE via our extended MVDM**

**(1) Confidence loss.** Previous studies typically interpret softmax-normalized logits as probability vectors summing to one. Based on this assumption, confidence loss in OE constrains the predictive distribution of auxiliary outliers, forcing it to approximate a uniform distribution, *i.e.*,

$$\begin{aligned} \mathcal{L}_{\text{OE}} = & \frac{1}{N} \sum_{j=1}^{N} - \log P_\theta(\hat{y} = y | X_j^{in}) + \\ & \frac{\beta}{M} \sum_{j=1}^{M} \sum_{i=1}^{k} - \frac{1}{k} \log P_\theta(\hat{y} = i | X_j^{out}) \end{aligned} \quad (4)$$

where $y$ is the one-hot label of ID sample $X_j^{in}$, $\beta$ is a loss weight for confidence-loss term and $\frac{1}{k}$ is the label on each category for OOD sample $X_j^{out}$.

**(2) Pitfalls of OE on logit calibration.** Our extended MVDM offers a logit calibration lens to revisit OE's underlying assumption. As shown in Figure 2 (a) left, OE assumes outliers only contain random noise. Consequently, OE with confidence loss penalizes models to output uniform softmax-normalized logits, and produce approximate logits for all




(a) Underlying **OOD data structure** in OE (left) and our MVOL (right). (b) Optimization objective on **logits** of OE and MVOL for outliers.

Figure 2: (a) Illustration of OOD data structure assumed in OE and our MVOL. (b) From our new assumption on out-of-distribution, OE blindly aligns the logits of all categories to the same level. Instead, MVOL adaptively aligns the logits of categories with associated minor ID features on outliers. The key insight is that a well-calibrated model should not display significantly distinct activation on these categories since minor ID features usually have small coefficients.

categories, as shown in Figure 2 (b) above. However, outliers might contain minor ID features that should not be ignored, as shown in Figure 2 (a) right that depicts a more realistic OOD data structure. Although unexpected high logits on outliers could be suppressed, confidence loss treats all categories equally, ignoring the inherent discrepancy between the categories only with noise (type I) and those with minor ID features (type II) in outliers. As a result, OE would overemphasize noise and underemphasize minor ID features when aligning the logits of all categories to the same level.

**Multi-view based learning objective.** Given the typical problem of OE, we introduce a new learning objective. In Figure 2 (b) below, MVOL adaptively aligns the logits of categories with associated minor ID features on outliers, *i.e.*,

$$\mathcal{L}_{\text{MVOL}}^{(t)} = \frac{1}{N}\sum_{j=1}^{N} -\log P_\theta(\hat{y}=y|X_j^{in}) \tag{5}$$
$$+ \frac{\beta}{M}\sum_{j=1}^{M}\sum_{i=1}^{k} -p_{j,i}^{(t)} \log P_\theta(\hat{y}=i|X_j^{out}),$$

$$\text{where} \quad p_{j,i}^{(t)} = \mathbf{min}(\mathbf{logit}_i(F^{(t)}, X_j^{out}), \epsilon), \tag{6}$$

$\mathbf{logit}_i(F,X) \overset{\text{def}}{=} \frac{e^{F_i(X)}}{\sum_{n\in[k]} e^{F_n(X)}}$. We take $S_{j,I}^{(t)} = \{i \in [k] | \mathbf{logit}_i(F^{(t)}, X_j^{out}) \leq \epsilon\}$, $S_{j,II}^{(t)} \overset{\text{def}}{=} \{i \in [k] | \mathbf{logit}_i(F^{(t)}, X_j^{out}) > \epsilon\}$ to represent those categories with smaller softmax-normalized values than threshold $\epsilon$ or not on outlier $X_j$. In Equation (6), if $\mathbf{logit}_i(F^{(t)}, X_j^{out})$ exceeds $\epsilon$, $p_{j,i}$ is set to $\epsilon$. Conversely, if $\mathbf{logit}_i(F^{(t)}, X_j^{out})$ is less than $\epsilon$, $p_{j,i}$ adopts $\mathbf{logit}_i(F^{(t)}, X_j^{out})$ directly. $p_{j,i}$ can be cast as the soft label adopted in Equation (5). The rationale behind our learning objective will be further explained through the parameter update mechanism, *i.e.*,

$$w_{i,r}^{(t+1)} = w_{i,r}^{(t)} - \eta\nabla_{w_{i,r}} L(F^{(t)}) - \frac{\eta^{'}}{M}\sum_{j=1}^{M}\Big[$$
$$\Big(\mathbf{logit}_i(F^{(t)}, X_j^{out})\sum_{n=1}^{k} p_{j,n}^{(t)} - p_{j,i}^{(t)}\Big)\nabla_{w_{i,r}} F_i^{(t)}(X_j^{out})\Big]. \tag{7}$$

where $L(F^{(t)}$ means a cross-entropy loss for ID data.

**(1) Detailed analysis.** Based on our data model, type II categories usually exhibit much larger logits than type I categories. **(i)** $S_{j,II}^{(t)}$ is generally associated with type II categories. By uniformly setting their labels to the threshold $\epsilon$, our learning objective can moderate the logits of these categories to ensure a smoother logit response, based on the insight that a well-calibrated model should not display significantly distinct activation values since minor ID features often have smaller coefficients. This allows our learning objective to adaptively adjust the logit values for type II categories rather than blindly suppressing them via aligning them to logits of type I categories. **(ii)** $S_{j,I}^{(t)}$ is often associated with type I categories. By setting $p_{j,i}\ i \in S_{j,I}^{(t)}$ as the softmax-normalized value itself, the gradient weight in Equation (5) can be written as a non-positive value $(\sum_{n=1}^{k} p_{j,n}^{(t)} - 1)p_{j,i}$, where the magnitude is proportional to $p_{j,i}$. Here, a lower $p_{j,i}$ leads to a smaller weight, thus contributing less to parameter updates. Thus, our $\mathcal{L}_{\text{MVOL}}$ effectively reduces the excessive focus on noise.

**(2) Adaptability to ID noise in wild datasets.** When auxiliay datasets contain ID data as noise, i.e., wild datasets (Yang et al. 2021; Katz-Samuels et al. 2022), eliminating all ID noise is labor-intensive. Benefiting from explicitly utilizing ID attributes in outliers, our learning objective demonstrates strong adaptability to ID data in wild datasets. In specific, when encountering ID data with ground truth label $y$ in wild datasets, models tend to produce a significantly larger logit for the category $y$ than the others. This reaction causes the coefficient $\sum_{n=1}^{k} p_{j,n}^{(t)}$ in Equation (7) to be small, reducing its impact on the logit of $y$, rather than blindly suppressing it. Thus, our learning objective can generalize to ID data in wild datasets. Relevant experiments are in Table 2.

# Experiments

## Experimental Setup

**Datasets. (1)** Following the common benchmarks in literature, we use CIFAR-10 and CIFAR-100 as ID datasets, and six diverse OOD test sets. The Tiny Images dataset used in (Hendrycks, Mazeika, and Dietterich 2019; Liu et al. 2020) for auxiliary outliers has been withdrawn due to its ethical wrong. Instead, we use 300K RandomImages, a cleaned subset of the Tiny Images dataset by (Hendrycks, Mazeika, and Dietterich 2019), as done in (Katz-Samuels et al. 2022). **(2)** For experiments with wild datasets, we partition CIFAR-10 into 2 halves: one provides ID training data, and the other

| | Category | Method | CIFAR-10 | | | CIFAR-100 | | |
|---|---|---|---|---|---|---|---|---|
| | | | FPR95 ↓ | AUROC ↑ | ID-Acc ↑ | FPR95 ↓ | AUROC ↑ | ID-Acc ↑ |
| Single Model Setting | Post Hoc | MSP | 56.29 ± 1.62 | 89.59 ± 0.63 | 94.27 ± 0.14 | 80.75 ± 0.81 | 74.70 ± 1.01 | <u>74.69</u> ± 0.21 |
| | | Energy | 41.20 ± 5.28 | 89.70 ± 1.93 | 94.27 ± 0.14 | 72.58 ± 1.78 | 79.01 ± 1.12 | 74.69 ± 0.21 |
| | | MaxLogit | 41.68 ± 4.99 | 89.69 ± 1.88 | 94.27 ± 0.14 | 73.21 ± 1.69 | 78.88 ± 1.11 | 74.69 ± 0.21 |
| | | ODIN | 41.75 ± 3.86 | 87.38 ± 2.41 | 94.27 ± 0.14 | 68.13 ± 1.83 | 79.36 ± 0.91 | 74.69 ± 0.21 |
| | | Mahalanobis | 23.96 ± 1.26 | 92.81 ± 0.32 | 94.27 ± 0.14 | 46.40 ± 3.73 | 87.44 ± 1.12 | 74.69 ± 0.21 |
| | | KNN | 30.89 ± 2.76 | 94.53 ± 0.44 | 94.27 ± 0.14 | 82.02 ± 2.58 | 75.84 ± 1.35 | 74.69 ± 0.21 |
| | | ASH | 40.03 ± 5.18 | 90.01 ± 1.70 | 94.26 ± 0.11 | 63.31 ± 1.91 | 79.35 ± 1.09 | 74.23 ± 0.31 |
| | Outlier Synthesis | VOS | 34.67 ± 5.01 | 91.54 ± 1.92 | **94.75** ± 0.17 | 70.17 ± 2.52 | 81.73 ± 1.78 | **75.94** ± 0.20 |
| | | ATOL | 12.86 ± 0.59 | 97.34 ± 0.07 | 93.89 ± 0.17 | 64.67 ± 1.73 | 80.17 ± 1.34 | 72.70 ± 0.17 |
| | Outlier Exposure | Energy w/Aux | 4.70 ± 0.50 | 97.77 ± 0.06 | 90.74 ± 0.24 | 52.43 ± 3.51 | 88.40 ± 1.16 | 62.13 ± 0.27 |
| | | OE | 4.25 ± 0.15 | 98.56 ± 0.07 | 94.47 ± 0.13 | 46.51 ± 3.65 | 89.78 ± 0.98 | 74.02 ± 0.04 |
| | | OE + MaxLogit | <u>4.12</u> ± 0.20 | <u>98.58</u> ± 0.07 | 94.47 ± 0.13 | <u>46.20</u> ± 3.53 | <u>90.59</u> ± 0.87 | 74.02 ± 0.04 |
| | | **MVOL** (ours) | **3.30** ± 0.19 | **98.70** ± 0.05 | <u>94.68</u> ± 0.09 | **42.96** ± 0.86 | **90.69** ± 0.26 | 74.29 ± 0.33 |
| nsemble Distillation Model Setting | Post Hoc | MSP | 55.82 ± 2.46 | 89.52 ± 0.53 | 94.36 ± 0.11 | 80.36 ± 0.72 | 74.72 ± 0.79 | <u>76.99</u> ± 0.15 |
| | | Energy | 38.23 ± 1.72 | 89.97 ± 0.57 | 94.36 ± 0.11 | 71.97 ± 1.02 | 79.66 ± 0.57 | 76.99 ± 0.15 |
| | | MaxLogit | 38.64 ± 2.00 | 89.94 ± 0.57 | 94.36 ± 0.11 | 72.71 ± 1.08 | 79.46 ± 0.60 | 76.99 ± 0.15 |
| | | ODIN | 40.46 ± 2.20 | 86.94 ± 1.07 | 94.36 ± 0.11 | 67.71 ± 0.61 | 78.71 ± 0.78 | 76.99 ± 0.15 |
| | | Mahalanobis | 23.06 ± 0.64 | 92.94 ± 0.19 | 94.36 ± 0.11 | <u>42.36</u> ± 1.93 | 88.39 ± 0.49 | 76.99 ± 0.15 |
| | | KNN | 41.98 ± 2.47 | 92.70 ± 0.55 | 94.36 ± 0.11 | 84.67 ± 2.72 | 73.67 ± 1.78 | 76.99 ± 0.15 |
| | | ASH | 36.50 ± 1.43 | 91.55 ± 0.55 | 94.23 ± 0.09 | 59.19 ± 1.75 | 80.20 ± 0.49 | 76.41 ± 0.26 |
| | Outlier Synthesis | VOS | 30.58 ± 4.34 | 92.23 ± 1.07 | **95.02** ± 0.11 | 72.01 ± 1.81 | 79.86 ± 1.90 | **77.18** ± 0.28 |
| | | ATOL | 28.14 ± 1.79 | 93.60 ± 0.43 | 94.19 ± 0.07 | 74.07 ± 0.98 | 77.82 ± 0.66 | 74.79 ± 0.09 |
| | Outlier Exposure | Energy w/Aux | 4.10 ± 0.24 | 98.07 ± 0.04 | 91.48 ± 0.19 | 52.81 ± 3.52 | 89.14 ± 0.81 | 68.27 ± 0.48 |
| | | OE | 3.95 ± 0.23 | 98.56 ± 0.07 | 94.67 ± 0.23 | 47.04 ± 0.73 | 89.35 ± 0.21 | 75.01 ± 0.13 |
| | | OE + MaxLogit | <u>3.61</u> ± 0.24 | **98.62** ± 0.06 | 94.67 ± 0.23 | 46.92 ± 0.75 | **90.79** ± 0.26 | 75.01 ± 0.13 |
| | | **MVOL** (ours) | **3.34** ± 0.20 | <u>98.61</u> ± 0.06 | <u>94.68</u> ± 0.20 | **36.62** ± 1.36 | <u>90.37</u> ± 0.43 | 76.27 ± 0.33 |

Table 1: Main results on out-of-distribution detection. The best result is in bold and the second is underlined. Values are averaged over five runs. Post Hoc-based methods use the same five pre-trained models but apply different scoring functions.

provides auxiliary noisy ID data. CIFAR-100 provides auxiliary outliers. The auxiliary ID and OOD data are mixed with varied noise levels $\alpha$, *i.e.*, 0, 0.05, 0.1, 0.3, 0.5, and the total images in auxiliary datasets are maintained at 50000. This setup allows simulating the scenarios, where auxiliary OOD datasets contain considerable noise yet most are still outliers.

**Evaluation Metrics.** We evaluate methods using common metrics, averaged on six OOD test sets, i.e., the false positive rate of declaring OOD examples as ID when 95% of ID data points are declared as ID (FPR95) and the area under the receiver operating characteristic curve (AUROC).

**OOD Detection Baselines.** We compare MVOL with comprehensive baselines. **(1)** Post-Hoc based: MSP (Hendrycks and Gimpel 2017), MaxLogit(Hendrycks et al. 2022), ODIN (Liang, Li, and Srikant 2018), Mahalanobis (Lee et al. 2018b), and Energy (Liu et al. 2020); **(2)** outlier synthesis-based: VOS (Du et al. 2022) and ATOL (Zheng et al. 2023), **(3)** outlier exposure-based: OE (Hendrycks, Mazeika, and Dietterich 2019) and Energy (Liu et al. 2020). Moreover, OE with MaxLogit as the OOD score, dubbed OE + MaxLogit, is our most relevant baseline. In wild-dataset settings, the SOTA WOODS is (Katz-Samuels et al. 2022) for reference.

**Training Details. (1)** We use the Wide ResNet with 40 layers and a widen factor of 2 for all experiments. **(2)** Involved two experimental settings: (i) *Single model*: single neural network is trained with one-hot labels and cross entropy loss. (ii) *Ensemble distillation model*: a single network is trained with the standard ensemble distillation algorithm in (Hinton, Vinyals, and Dean 2015), where soft labels are generated by ensemble models.

## Results

**Main results on OOD detection.** In this experiment, Outlier Exposure-based methods utilize auxiliary OOD datasets. Through comparisons in Table 1, we have four main findings. **(1)** OE + MaxLogit outperforms OE (MSP as OOD scoring function), usually serving as the best baseline. This supports our analysis with our extended MVDM, *i.e.*, OE performs logit calibration on outliers. **(2)** Our MVOL obtains an overall stronger performance than other methods. In the ensemble distillation model setting, MVOL can reduce FPR95 by 10.3% on CIFAR-100, with a slight decrease of 0.04% in AUROC. **(3)** With CIFAR-100 as an ID dataset, compared to MVOL in the single model setting, MVOL in the ensemble distillation model setting can reduce FPR95% by 6.34%. This result supports the comparison of Theorem 1 and 2. It also indicates the superiority of MaxLogit when models learns all features. There is a small improvement on CIFAR-10. The reason could be that a single model learns

| Noise | Method | Single Model Setting | | | Ensemble Distillation Model Setting | | |
|---|---|---|---|---|---|---|---|
| | | FPR95 ↓ | AUROC ↑ | ID-Acc ↑ | FPR95 ↓ | AUROC ↑ | ID-Acc ↑ |
| - | MaxLogit | 47.39 ± 2.93 | 89.81 ± 0.55 | 91.38 ± 0.24 | 40.59 ± 2.50 | 90.21 ± 0.71 | 91.94 ± 0.09 |
| $\alpha = 0$ | WOODS | 21.97 ± 1.83 | 96.02 ± 0.32 | 91.20 ± 0.22 | 51.50 ± 3.99 | 84.85 ± 0.99 | 89.79 ± 0.19 |
| | OE + MaxLogit | 18.04 ± 1.34 | **96.57** ± 0.15 | **92.24** ± 0.08 | 18.86 ± 2.10 | 96.12 ± 0.31 | **92.32** ± 0.15 |
| | **MVOL** (ours) | **17.34** ± 2.86 | 96.21 ± 0.30 | 91.71 ± 0.08 | **12.96** ± 0.95 | **96.45** ± 0.14 | 91.96 ± 0.13 |
| $\alpha = 0.05$ | WOODS | 22.04 ± 2.36 | 96.02 ± 0.34 | 91.27 ± 0.16 | 51.49 ± 3.97 | 84.86 ± 0.99 | 89.79 ± 0.19 |
| | OE + MaxLogit | 22.11 ± 1.26 | 95.64 ± 0.26 | 91.29 ± 0.20 | 23.11 ± 3.90 | 95.67 ± 0.48 | 91.51 ± 0.18 |
| | **MVOL** (ours) | **19.55** ± 1.04 | **96.22** ± 0.16 | **91.75** ± 0.23 | **12.87** ± 1.11 | **96.49** ± 0.11 | **91.74** ± 0.30 |
| $\alpha = 0.1$ | WOODS | 22.38 ± 2.30 | 95.98 ± 0.35 | 91.27 ± 0.21 | 51.59 ± 4.03 | 84.85 ± 0.98 | 89.81 ± 0.19 |
| | OE + MaxLogit | 25.49 ± 1.60 | 94.97 ± 0.30 | 90.92 ± 0.31 | 26.90 ± 3.04 | 95.24 ± 0.41 | 91.01 ± 0.22 |
| | **MVOL** (ours) | **18.05** ± 1.58 | **96.16** ± 0.20 | **91.55** ± 0.31 | **13.96** ± 0.80 | **96.07** ± 0.15 | **91.64** ± 0.28 |
| $\alpha = 0.3$ | WOODS | 22.03 ± 2.45 | 95.99 ± 0.40 | 91.24 ± 0.19 | 51.58 ± 4.03 | 84.86 ± 0.99 | 89.80 ± 0.21 |
| | OE + MaxLogit | 37.67 ± 3.85 | 92.64 ± 0.44 | 89.04 ± 0.30 | 51.20 ± 6.17 | 91.83 ± 0.57 | 88.60 ± 0.13 |
| | **MVOL** (ours) | **20.71** ± 1.86 | **95.94** ± 0.32 | **91.24** ± 0.16 | **13.79** ± 0.93 | **96.43** ± 0.16 | **91.76** ± 0.07 |
| $\alpha = 0.5$ | WOODS | **22.31** ± 3.05 | **95.93** ± 0.50 | **91.28** ± 0.12 | 51.49 ± 3.96 | 84.86 ± 0.99 | 89.80 ± 0.20 |
| | OE + MaxLogit | 45.14 ± 2.94 | 90.22 ± 0.63 | 88.04 ± 0.18 | 54.23 ± 3.77 | 89.93 ± 0.21 | 86.80 ± 0.21 |
| | **MVOL** (ours) | 25.53 ± 1.23 | 95.23 ± 0.18 | 90.79 ± 0.21 | **14.85** ± 1.86 | **96.44** ± 0.32 | **91.81** ± 0.19 |

Table 2: Main results on OOD detection with wild datasets (a larger $\alpha$ means more ID noise). WOODS is for reference.

more features on CIFAR-10 than CIFAR-100, as CIFAR-10 has more samples per class (Allen-Zhu and Li 2023). **(4)** Across both settings on two ID datasets, MVOL often degrades less ID accuracy than other OE-based methods.

Furthermore, we analyze that MVOL can utilize ID attributes in outliers as follows. **(1)** In Figure 3 (a), compared to OE + MaxLogit, MVOL not only preserves ID features on OOD data but also mitigates overemphasizing noise. It verified our key insight shown in Figure 2 (b). **(2)** In Table 1, we observe that MVOL generally shows greater superiority in FPR95 than AUROC. Based on this, we further analyze MVOL focus on ID attributes in outliers. In Figure 3 (b), compared to OE + MaxLogit, MVOL achieves a lower FPR at a high TPR, e.g., 0.95. This suggests that MVOL has better discriminability of ID against OOD, as a high TPR means a high recall of ID samples. In a unified perspective on ID and OOD, logit responses to OOD data would affect ID data. This discriminability could due to MVOL not blindly suppressing logit responses to ID attributes in outliers. Instead, MVOL adaptively calibrates these logits. As a result, it reserves logit responses to ID data and improves discriminability with MaxLogit. In addition, the smaller FPR of OE + MaxLogit, when the TPR is low, could be due to overfitting a subset of ID data against OOD, increasing AUROC.

**Main results on OOD detection with wild datasets.** This experiment simulates a real scenario where collected auxiliary OOD datasets contain ID data as noise. We use OE + MaxLogit as the baseline for comparison. Because compared to it head-to-head, our MVOL has a new multi-view-based learning objective ($\mathcal{L}_{\text{MVOL}}$) specially designed to tackle intrinsic ID attributes in outliers. We use wild datasets to verify whether $\mathcal{L}_{\text{MVOL}}$ could generalize well to noisy ID data in them. Furthermore, we use WOODS as the SOTA baseline in wild-dataset settings only for reference, as it is specifically designed to handle noisy ID data in wild dataset.

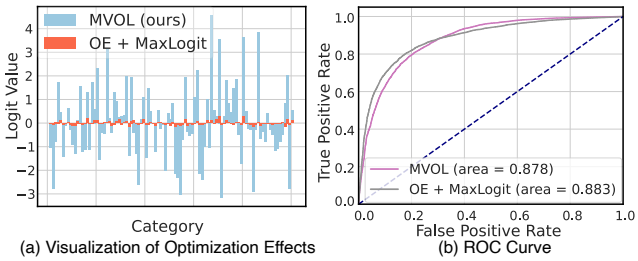



Figure 3: MVOL and OE + MaxLogit with CIFAR-100 as ID dataset. (a) Logits of a training OOD sample, (b) ROC curve with Textures as the test OOD dataset.

We have two main findings in Table 2. **(1)** MVOL shows advantages using wild datasets in both settings. For example, in single model setting, $\mathcal{L}_{\text{MVOL}}$ reduces FPR95 by 0.7%, 2.56%, 7.44%, 16.96%, and 19.61% as alpha increases from 0 to 0.5, showing increasingly clear advantages over OE + MaxLogit, which confirms its generalizability. **(2)** MVOL is comparable with the customized WOODS when alpha=0.5 in single model setting. WOODS degrades in ensemble distillation model setting, even worse than MaxLogit (without auxiliary data). The reason could be its inefficacy in benefiting from distilled knowledge of ensemble models.

## Conclusion

This paper explores the intrinsic ID attributes in outliers. We propose a structured multi-view-based out-of-distribution learning framework to handle these attributes rationally. We provide theoretical analysis on it. Extensive experiments on auxiliary OOD datasets and even wild datasets show its efficacy. Through our unified perspective on ID and OOD, we probably calibrate logits even without auxiliary outliers, reducing computation costs. We leave it as future works.

## Acknowledgments

This work is supported by the National Natural Science Foundation of China (NSFC) under Grant No. 62476049.

## References

Allen-Zhu, Z.; and Li, Y. 2023. Towards Understanding Ensemble, Knowledge Distillation and Self-Distillation in Deep Learning. In *The Eleventh International Conference on Learning Representations*.

Bai, H.; Canal, G.; Du, X.; Kwon, J.; Nowak, R. D.; and Li, Y. 2023. Feed two birds with one scone: Exploiting wild data for both out-of-distribution generalization and detection. In *International Conference on Machine Learning*, 1454–1471.

Chen, J.; Li, Y.; Wu, X.; Liang, Y.; and Jha, S. 2021. ATOM: Robustifying Out-of-Distribution Detection Using Outlier Mining. In *Proceedings of European Conference on Machine Learning and Principles and Practice of Knowledge Discovery in Databases (ECML-PKDD)*, 430–445.

Djurisic, A.; Bozanic, N.; Ashok, A.; and Liu, R. 2023. Extremely Simple Activation Shaping for Out-of-Distribution Detection. In *The Eleventh International Conference on Learning Representations*.

Du, X.; Sun, Y.; Zhu, J.; and Li, Y. 2023. Dream the Impossible: Outlier Imagination with Diffusion Models. In *Advances in Neural Information Processing Systems*, volume 36, 60878–60901. Curran Associates, Inc.

Du, X.; Wang, Z.; Cai, M.; and Li, Y. 2022. VOS: Learning What You Don't Know by Virtual Outlier Synthesis. In *International Conference on Learning Representations*.

Hendrycks, D.; Basart, S.; Mazeika, M.; Zou, A.; Kwon, J.; Mostajabi, M.; Steinhardt, J.; and Song, D. 2022. Scaling Out-of-Distribution Detection for Real-World Settings. In *Proceedings of the 39th International Conference on Machine Learning*, volume 162, 8759–8773.

Hendrycks, D.; and Gimpel, K. 2017. A Baseline for Detecting Misclassified and Out-of-Distribution Examples in Neural Networks. In *International Conference on Learning Representations*.

Hendrycks, D.; Mazeika, M.; and Dietterich, T. 2019. Deep Anomaly Detection with Outlier Exposure. In *International Conference on Learning Representations*.

Hinton, G.; Vinyals, O.; and Dean, J. 2015. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*.

Katz-Samuels, J.; Nakhleh, J. B.; Nowak, R.; and Li, Y. 2022. Training OOD Detectors in their Natural Habitats. In *Proceedings of the 39th International Conference on Machine Learning*, volume 162, 10848–10865.

Lee, K.; Lee, H.; Lee, K.; and Shin, J. 2018a. Training Confidence-calibrated Classifiers for Detecting Out-of-Distribution Samples. In *International Conference on Learning Representations*.

Lee, K.; Lee, K.; Lee, H.; and Shin, J. 2018b. A Simple Unified Framework for Detecting Out-of-Distribution Samples and Adversarial Attacks. In *Advances in Neural Information Processing Systems*, volume 31.

Liang, S.; Li, Y.; and Srikant, R. 2018. Enhancing The Reliability of Out-of-distribution Image Detection in Neural Networks. In *International Conference on Learning Representations*.

Liu, W.; Wang, X.; Owens, J.; and Li, Y. 2020. Energy-based out-of-distribution detection. *Advances in neural information processing systems*, 33: 21464–21475.

Ming, Y.; Fan, Y.; and Li, Y. 2022. POEM: Out-of-Distribution Detection with Posterior Sampling. In *Proceedings of International Conference on Machine Learning (ICML)*, 15650–15665.

Nguyen, A.; Yosinski, J.; and Clune, J. 2015. Deep Neural Networks Are Easily Fooled: High Confidence Predictions for Unrecognizable Images. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Tao, L.; Du, X.; Zhu, J.; and Li, Y. 2023. Non-parametric Outlier Synthesis. In *The Eleventh International Conference on Learning Representations*.

Vaze, S.; Han, K.; Vedaldi, A.; and Zisserman, A. 2022. Open-Set Recognition: A Good Closed-Set Classifier is All You Need. In *International Conference on Learning Representations*.

Wei, H.; Xie, R.; Cheng, H.; Feng, L.; An, B.; and Li, Y. 2022. Mitigating Neural Network Overconfidence with Logit Normalization. In *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, 23631–23644.

Yang, J.; Wang, H.; Feng, L.; Yan, X.; Zheng, H.; Zhang, W.; and Liu, Z. 2021. Semantically Coherent Out-of-Distribution Detection. In *2021 IEEE/CVF International Conference on Computer Vision, ICCV 2021, Montreal, QC, Canada, October 10-17, 2021*, 8281–8289.

Zheng, H.; Wang, Q.; Fang, Z.; Xia, X.; Liu, F.; Liu, T.; and Han, B. 2023. Out-of-distribution detection learning with unreliable out-of-distribution sources. In *Advances in Neural Information Processing Systems*, volume 36, 72110–72123.

Zhu, J.; Geng, Y.; Yao, J.; Liu, T.; Niu, G.; Sugiyama, M.; and Han, B. 2023. Diversified Outlier Exposure for Out-of-Distribution Detection via Informative Extrapolation. In *Advances in Neural Information Processing Systems*, volume 36, 22702–22734.