Image-based Outlier Synthesis With Training Data

Sudarshan Regmi
Department of Computer Science
Dartmouth College
sudarshan.regmi.gr@dartmouth.edu

Abstract

Out-of-distribution (OOD) detection is critical to ensure the safe deployment of deep learning models in critical applications. Deep learning models can often misidentify OOD samples as in-distribution (ID) samples. This vulnerability worsens in the presence of spurious correlation in the training set. Likewise, in fine-grained classification settings, detection of fine-grained OOD samples becomes inherently challenging due to their high similarity to ID samples. However, current research on OOD detection has focused instead largely on relatively easier (conventional) cases. Even the few recent works addressing these challenging cases rely on carefully curated or synthesized outliers, ultimately requiring external data. This motivates our central research question: "Can we innovate OOD detection training framework for fine-grained and spurious settings ***without requiring any external data at all?****" In this work, we present a unified* ***A****pproach to* ***S****purious, fine-grained, and* ***C****onventional* ***OOD D****etection (****ASCOOD****) that eliminates the reliance on external data. First, we synthesize virtual outliers from ID data by approximating the destruction of invariant features. Specifically, we propose to add gradient attribution values to ID inputs to disrupt invariant features while amplifying true-class logit, thereby synthesizing challenging near-manifold virtual outliers. Then, we simultaneously incentivize ID classification and predictive uncertainty towards virtual outliers. For this, we further propose to leverage standardized features with z-score normalization. ASCOOD effectively mitigates impact of spurious correlations and encourages capturing fine-grained attributes. Extensive experiments across* ***7*** *datasets and comparisons with* ***30+*** *methods demonstrate merit of ASCOOD in spurious, fine-grained and conventional settings.*

1. Introduction

Deploying deep learning models, trained under the *closed-world* assumption ($\mathbb{D}_{\text{train}} = \mathbb{D}_{\text{test}}$), often becomes challenging in real-world scenarios as they frequently encounter OOD inputs. OOD inputs should be accurately flagged as they lie beyond the training distribution. Such identification of OOD inputs becomes challenging if models rely on spurious features that do not generalize beyond the training distribution [56]. For instance, a medical diagnosis model might erroneously rely on spurious features, such as image artifacts, leading it to incorrectly classify any image containing such artifacts as ID sample. Similarly, in fine-grained scenarios [10, 28, 96, 107] like species classification, novel species visually similar to known ones can easily be misidentified as ID species. Proper consideration of these scenarios while ensuring high ID accuracy is essential for safe deployment of deep learning models.

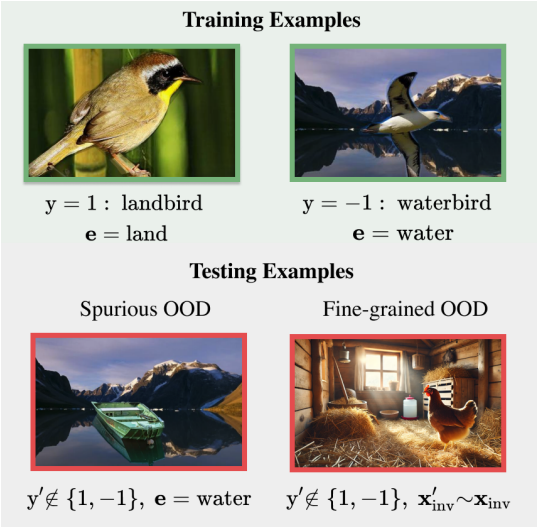


Figure 1. In Waterbirds dataset [69], label $y \in \{\text{waterbird}, \text{landbird}\}$ is correlated with environmental feature $\mathbf{e} \in \{\text{water}, \text{land}\}$. Spurious OOD retains environmental feature $\mathbf{e}$ (water) while fine-grained OOD has its invariant feature similar to ID invariant feature ($\mathbf{x}'_{\text{inv}} \sim \mathbf{x}_{\text{inv}}$). Both present significant challenges for OOD detection.

Images generally consist of both *invariant* and *environmental* features. As shown in Figure 1, when the correlations between environmental features (land and water) and

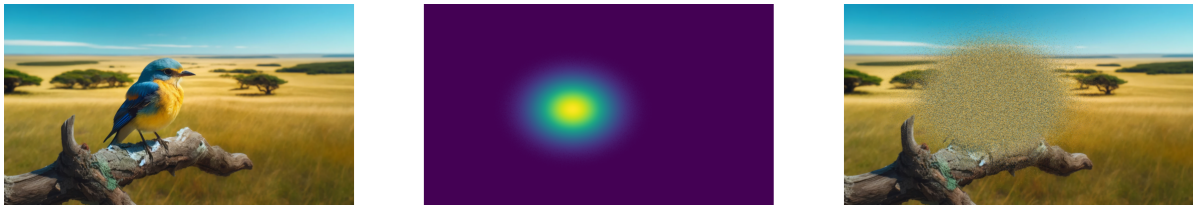


Figure 2. Motivating example of the outlier synthesis pipeline. **Left:** An image $\mathbf{x} = \psi(\mathbf{x}_{\text{inv}}, \mathbf{e}) \in \mathcal{X}$ from the in-distribution dataset $\mathbb{D}_{\text{in}}$ is shown. **Middle:** A 2D distribution $\mathcal{G}_{\text{oracle}}$ is shown which signifies the presence of invariant feature in a smaller region of the image $\mathbf{x}$. **Right:** Corresponding outlier $\mathbf{x}'$ is shown, which is formed by destroying the invariant feature $\mathbf{x}_{\text{inv}}$ of $\mathbf{x}$ through a perturbation function $\mathcal{P}_F$, having access to $\mathcal{G}_{\text{oracle}}$. *Can we synthesize similar virtual outlier $\mathbf{x}'$ without the access of $\mathcal{G}_{\text{oracle}}$?*

corresponding target labels (landbird, waterbird) are high, neural networks can rely on spurious features to achieve high classification performance [7, 69]. It can cause the model to incorrectly make high-confidence predictions for OOD samples with similar environmental features but different semantic content. Moreover, in fine-grained classification settings, the degree of distinction between ID and OOD samples may be as subtle as that between different ID classes. As illustrated in Figure 1, fine-grained OOD for the Waterbirds ID dataset may be "hen", which differs from ID samples based on subtle fine-grained attributes (Also, see A.2). Moreover, the overlap of high-level feature sets between fine-grained OOD and ID data complicates the detection of the former. Real-world scenarios frequently involve either spurious or fine-grained settings as ID and OOD are often captured under similar conditions during deployment, highlighting the importance of study under such settings.

A significant majority of OOD detection studies, including recent ones [18, 19, 33, 44, 46, 97], restrict their studies to conventional cases. While few works [64, 72, 80, 102] study fine-grained OOD detection, they often require the curation of diverse outliers non-overlapping with ID data [4, 32, 98, 109]. Some recent works [2, 9, 18, 39, 46] use foundation models to synthesize the outliers in image space. Such approach can be computationally intensive, often requiring multiple steps and careful prompting to curate outliers. The reliance on domain knowledge of foundation model limits its applicability in highly novel scenarios.

On the other hand, Ming *et al*. [56] and Zhang *et al*. [104] have explored the detrimental effect of spurious correlation on OOD detection, but studies addressing this issue (with virtual outliers) leveraging only training samples remain relatively scarce. A few notable works such as Kirby [34] and BackMix [87] propose to use background image features utilizing inpainting procedure while OEST [85] utilizes explicit data augmentations. In this work however, we take a more direct simplistic approach – we add gradient attribution values to ID inputs to disrupt invariant features while amplifying true-class logit, thereby synthesizing challenging near-manifold virtual outliers.

To summarize, in this work, we propose a unified *Approach to **S**purious, fine-grained and **C**onventional OOD Detection* (dubbed **ASCOOD**). ASCOOD consists of: ① outlier synthesis pipeline and ② virtual outlier exposure (OE) training pipeline. To synthesize virtual outliers, we perturb invariant features while preserving environmental features. We identify invariant features with the pixel attribution method using the model being learned. Second, we formulate a joint training objective that incentivizes the ID classification and the predictive uncertainty toward the synthesized outliers. To facilitate the joint objective, we employ constrained optimization by leveraging standardized feature representation. Our contributions are:

- We propose a novel OE training approach leveraging standardized feature representation, along with an improved variant of posthoc method ODIN [45].
- To the best of our knowledge, we are the first to empirically demonstrate that adding gradient attribution values to ID samples synthesizes effective outliers, whereas subtracting these values does not. We also introduce invariant pixel shuffling as a strong outlier synthesis baseline.
- We empirically reveal superiority of z-score over L_2 normalization in feature representation for training the OOD detection model.

2. Preliminaries

Background: We consider supervised multi-class classification setup. Let $\mathcal{X}_{\text{inv}} \in \mathbb{R}^v$ denote invariant image space, where each invariant feature $\mathbf{x}_{\text{inv}} \in \mathcal{X}_{\text{inv}}$ is essential for class recognition. Let $\mathcal{Y} = \{1, 2, \ldots, C\}$ be label space consisting of C predefined classes with each label $y \in \mathcal{Y}$ associated with an invariant feature $\mathbf{x}_{\text{inv}}$. Let $\mathbf{y}$ be the one-hot vector of y. Let $\mathcal{E} \in \mathbb{R}^t$ denote environment space comprising o distinct environments $\{e_1, e_2, \ldots, e_o\}$. Input space $\mathcal{X} \in \mathbb{R}^{v+t}$ is defined such that each input $\mathbf{x} \in \mathcal{X}$ is a function ψ of $\mathbf{x}_{\text{inv}} \in \mathcal{X}_{\text{inv}}$ and $\mathbf{e} \in \mathcal{E}$, i.e., $\mathbf{x} := \psi(\mathbf{x}_{\text{inv}}, \mathbf{e})$, with $\mathbf{e}$ providing non-essential contextual cues. The training dataset $\mathbb{D}_{\text{train}} = \{(\mathbf{x}, \mathbf{y})_i \mid i = 1, 2, \ldots, N\}$ consists of N i.i.d. samples from distribution $P(\mathcal{X}, \mathcal{Y})$. A feature extractor $\phi_\gamma : \mathcal{X} \to \mathbb{R}^m$ maps input $\mathbf{x} \in \mathcal{X}$ to a feature $\mathbf{h} \in \mathcal{H}$ in feature space $\mathcal{H} \in \mathbb{R}^m$, i.e., $\mathbf{h} := \phi_\gamma(\mathbf{x})$. A clas-

sifier $f_\theta : \mathbb{R}^m \to \mathbb{R}^C$ assigns logits $\mathbf{z} \in \mathbb{R}^C$ to $\mathbf{h}$, which are transformed into probabilities $\mathbf{p} = \rho(\mathbf{z}) \in \mathbb{R}^C$ using softmax function: $\rho(\mathbf{z})_j = \frac{\exp(z_j)}{\sum_{l=1}^{C} \exp(z_l)}, \forall j \in [1, C]$. The classification model $g = f_\theta \circ \phi_\gamma$ is traditionally optimized under the *closed-world* assumption with empirical risk minimization using $\mathcal{L}$ loss function.: $\min_{\phi_\gamma, f_\theta} \mathcal{L}(\rho(f_\theta(\phi_\gamma(\mathbf{x}))), \mathbf{y})$.

OOD detection: The deployment of model g in open world (test distribution $\mathbb{D}_{\text{test}} = \{\mathbb{D}_{\text{train}}, \mathbb{D}_{\text{out}}\} = \{\mathbb{D}_{\text{in}}, \mathbb{D}_{\text{out}}\}$) violates the closed-world assumption ($\mathbb{D}_{\text{test}} = \mathbb{D}_{\text{train}}$), where $\mathbb{D}_{\text{out}}$ is OOD. OOD input $\mathbf{x}' \in \mathbb{D}_{\text{out}}$ should be correctly identified to ensure the safe operation of model g. This is generally achieved through a scoring function $s : \mathbb{R}^m \to \mathbb{R}$ (possibly incorporating f_θ), that quantifies the alignment of input $\mathbf{x}_{\text{test}}$ with $\mathbb{D}_{\text{in}}$ via the score $s(\phi_\gamma(\mathbf{x}_{\text{test}}))$. Specifically, if $s(\phi_\gamma(\mathbf{x}_{\text{test}})) \geq \beta$, it indicates $\mathbf{x}_{\text{test}} \in \mathbb{D}_{\text{in}}$. Conversely, if $s(\phi_\gamma(\mathbf{x}_{\text{test}})) < \beta$, it indicates $\mathbf{x}_{\text{test}} \in \mathbb{D}_{\text{out}}$. Here, β represents a threshold chosen to have a higher true positive rate (e.g., 95%) over the input space $\mathcal{X}$. If $\mathcal{E}' \in \mathbb{R}^t$ and $\mathcal{X}'_{\text{inv}} \in \mathbb{R}^v$ represent another environment and invariant input space respectively such that $\mathcal{X}_{\text{inv}} \cap \mathcal{X}'_{\text{inv}} = \emptyset$ and $\mathcal{E} \cap \mathcal{E}' = \emptyset$, we can formalize three kinds of OOD inputs: an input $\mathbf{x}_{\text{test}}$ is known as *conventional OOD* if $\mathbf{x}_{\text{test}} = \psi(\mathbf{x}'_{\text{inv}}, \mathbf{e}')$. It is known as *spurious OOD* if $\mathbf{x}_{\text{test}} = \psi(\mathbf{x}'_{\text{inv}}, \mathbf{e})$. In either case, $\mathbf{x}_{\text{test}}$ can be *fine-grained OOD* if $\mathbf{x}'_{\text{inv}} \sim \mathbf{x}_{\text{inv}}$.

3. Method

In this section, we motivate our method with an example and then formulate our learning framework based on this motivation. We subsequently detail the outlier synthesis and virtual outlier exposure training.

Motivation: As depicted in Figure 2 (left), we analyze an image $\mathbf{x} = \psi(\mathbf{x}_{\text{inv}}, \mathbf{e}) \in \mathbb{D}_{\text{in}}$ consisting of invariant feature (bird) $\mathbf{x}_{\text{inv}}$ and environmental feature (land) $\mathbf{e}$. Only a smaller portion contains the invariant feature $\mathbf{x}_{\text{inv}}$ necessary for class recognition, while the remainder comprises non-essential environmental features $\mathbf{e}$. *Can we synthesize challenging outlier $\mathbf{x}'$ from $\mathbf{x}$ by perturbing $\mathbf{x}_{inv}$ while retaining $\mathbf{e}$?* Let $\mathcal{G}_{\text{oracle}}$ denote an oracle 2D distribution indicating the presence of invariant feature $\mathbf{x}_{\text{inv}}$ in $\mathbf{x}$. Consider a transformation $\psi^{-1}_{\mathcal{G}_{\text{oracle}}}$, with access to $\mathcal{G}_{\text{oracle}}$, that decomposes $\mathbf{x}$ i.e. $\psi^{-1}_{\mathcal{G}_{\text{oracle}}}(\mathbf{x}) \to [\mathbf{x}_{\text{inv}}, \mathbf{e}]$. Consider a perturbation function $\mathcal{P}_F$ that disrupts the semantics of $\mathbf{x}_{\text{inv}}$, yielding $\mathbf{e}'$ such that $\mathcal{P}_F([\mathbf{x}_{\text{inv}}, \mathbf{e}]) = [\mathbf{e}', \mathbf{e}]$. Using these transformations, we can synthesize an outlier $\mathbf{x}' = \psi(\mathcal{P}_F(\psi^{-1}_{\mathcal{G}_{\text{oracle}}}(\mathbf{x})))$. *In the absence of $\mathcal{G}_{oracle}$, can we approximate it with $\mathcal{G}$ for each $\mathbf{x} \in \mathcal{X}$ to synthesize outlier $\mathbf{x}' \in \mathbb{D}_{out}$?* Training the network to enhance predictive uncertainty towards these challenging outliers improves the model's uncertainty towards OOD.

Learning framework: With the assumption of access to synthesized virtual $\mathbb{D}_{\text{out}}$, our learning framework is designed to optimize the parameters θ (of f_θ) and γ (of ϕ_γ) of a classification model g, simultaneously focusing on ID classification accuracy and uncertainty on OOD inputs. We define the total loss function, $\mathcal{L}_{\text{total}}$ as:

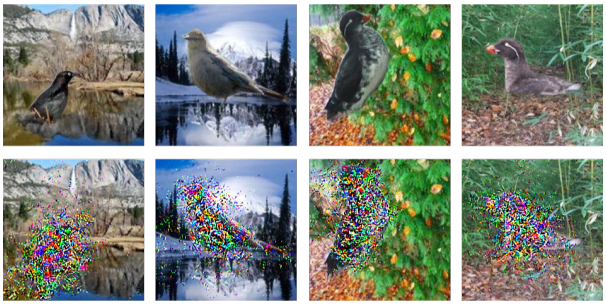


Figure 3. **Top row**: In-distribution images from the Waterbirds dataset. The first two images show waterbirds in water backgrounds, while the last two show landbirds in land backgrounds. **Bottom row**: Synthesized virtual outliers corresponding to the images in the top row at the latter stage of training.

$$\to \arg\min_{\theta,\gamma} \underbrace{\mathcal{L}_{\text{ID}}\left(f_\theta\left(\phi_\gamma(\mathbb{D}_{\text{in}})\right)\right)}_{\text{ID classification error}} + \underbrace{\mathcal{L}_{\text{OOD}}\left(f_\theta\left(\phi_\gamma(\mathbb{D}_{\text{out}})\right)\right)}_{\text{Uncerntainty error}} \tag{1}$$

We use cross-entropy loss $\mathcal{L}_{\text{CE}}$ for ID classification loss $\mathcal{L}_{\text{ID}}$ and KL divergence loss $\mathcal{L}_{\text{KL}}$ between virtual $\mathbb{D}_{\text{out}}$ and uniform distribution $\mathcal{U}$ for uncertainty loss $\mathcal{L}_{\text{OOD}}$.

3.1. Image-based Outlier Synthesis

We synthesize virtual outliers from input space $\mathcal{X}$ by approximately perturbing the invariant features $\mathbf{x}_{\text{inv}}$ while preserving the environmental features $\mathbf{e}$ of image $\mathbf{x}$. In the interpretability literature, several methods [3, 54, 73, 86, 89] have been proposed to compute saliency map that quantifies the importance of each pixel. A straightforward approach to computing it involves calculating derivative (i.e. gradient) $\mathbf{G}$ of the logit value of true class ($\mathbf{z}_c$) with respect to the input image $\mathbf{x}$:

$$\mathbf{G} = \frac{\partial \mathbf{z}_c}{\partial \mathbf{x}} \tag{2}$$

For an input $\mathbf{x}' = \mathbf{x} + \alpha \cdot \mathbf{G}$, the model g exhibits an increase in logit value of true class compared to the original input $\mathbf{x}$. Since input $\mathbf{x}'$ (with sufficiently high α) has its invariant features destroyed (rendering it an outlier), the model should ideally express uncertainty. On the other hand, an increase in the logit value of the true class (roughly speaking) suggests that $\mathbf{x}'$ can serve as a challenging outlier.

We observe similar empirical effects using gradients of either logits or softmax probabilities for outlier synthesis (see Sec. D.3). Since $\mathbf{G}$ assigns larger magnitudes to invariant pixels and smaller ones to environmental pixels, adding $\mathbf{G}$ to $\mathbf{x}$ disproportionately degrades invari-

ant features while minimally impacting environmental features. *Consequently, it effectively perturbs invariant features while preserving environmental features.* $\mathbf{G}$ can be sparsified by masking out the low-magnitude regions. Consider an image $\mathbf{x}$ consisting of $p_{\text{inv}}\%$ of pixels which are invariant pixels. Let $|\mathbf{G}|^{(100-p_{\text{inv}})\%}$ denote the $(100-p_{\text{inv}})^{\text{th}}$ percentile of $|\mathbf{G}|$. The gradient $\mathbf{G}$ with suppressed environment features can be expressed as:

$$\mathbf{G}_{\text{inv}}^{j} = \begin{cases} \mathbf{G}^{j}, & \text{if } |\mathbf{G}|^{j} \geq |\mathbf{G}|^{(100-p_{\text{inv}})\%} \\ 0, & \text{if } |\mathbf{G}|^{j} < |\mathbf{G}|^{(100-p_{\text{inv}})\%} \end{cases}$$

We compute $\mathbf{G}$ with the model being learned. In highly spurious settings, using $\mathbf{x}' = \mathbf{x} + \alpha \cdot \mathbf{G}_{\text{inv}}$ better preserves environmental features. The examples of synthesized outliers depicted in Figure 3 indeed show invariant features of the images being altered. Inspired by such perturbation, we propose improved variant of ODIN [45], **invariant-ODIN (i-ODIN)** (See Sec. B).

Additionally, we also propose a novel (to the best of our knowledge) way of synthesizing virtual outlier by shuffling invariant pixels $\mathbf{x}_{\text{inv}}$ of ID sample (See Sec. D). If `shuffle` denotes pixel-shuffling operation, virtual outliers could be synthesized as:

$$\mathbf{x}' = \psi(\ \texttt{shuffle}(\mathbf{x}_{\text{inv}}),\ \mathbf{e}\ )$$

3.2. Virtual Outlier Exposure (OE) Training:

We propose to train model g by simultaneously optimizing ID classification and predictive uncertainty towards the outliers with the learning framework of Equation 1.

Proposition 1 *The derivative of $\mathcal{L}_{total} = \mathcal{L}_{CE} + \mathcal{L}_{KL}$ w.r.t k^{th} logit is $(\boldsymbol{p}_k - \boldsymbol{y}_k) + (\boldsymbol{p}'_k - 1/C)$.*

Proof. The cross-entropy loss $\mathcal{L}_{\text{CE}}$ is given by:

$$\mathcal{L}_{\text{CE}} = -\sum_{l=1}^{C} \mathbf{y}_l \log \mathbf{p}_l, \quad \mathbf{p}_l = \rho(\mathbf{z}_l) = \frac{\exp(\mathbf{z}_l)}{\sum_{r=1}^{C} \exp(\mathbf{z}_r)}.$$

To compute $\frac{\partial \mathcal{L}_{\text{CE}}}{\partial \mathbf{z}_k}$, we proceed by substituting the $\mathbf{p}_l$ in $\mathcal{L}_{\text{CE}}$ and performing log expansion.

$$\mathcal{L}_{\text{CE}} = -\sum_{l=1}^{C} \mathbf{y}_l \mathbf{z}_l + \log\left(\sum_{r=1}^{C} \exp(\mathbf{z}_r)\right)$$

$$\text{Hence, } \frac{\partial \mathcal{L}_{\text{CE}}}{\partial \mathbf{z}_k} = (\mathbf{p}_k - \mathbf{y}_k)$$

The Kullback-Leibler divergence loss $\mathcal{L}_{\text{KL}}$ is given by:

$$\mathcal{L}_{\text{KL}} = \sum_{l=1}^{C} \mathbf{p}'_l \log(\mathbf{p}'_l) - \sum_{l=1}^{C} \mathbf{p}'_l \log\left(\frac{1}{C}\right)$$

$$\text{Hence, } \frac{\partial \mathcal{L}_{\text{KL}}}{\partial \mathbf{z}'_k} = (\mathbf{p}'_k - 1/C)$$

$$\text{So, } \frac{\partial \mathcal{L}_{\text{CE}}}{\partial \mathbf{z}_k} + \frac{\partial \mathcal{L}_{\text{KL}}}{\partial \mathbf{z}'_k} = (\mathbf{p}_k - \mathbf{y}_k) + (\mathbf{p}'_k - 1/C)$$

□

During the initial phase of training, model g lacks a comprehensive understanding of ID features. As we rely on the model for outlier synthesis, it may fail to synthesize true outliers ($\mathbf{x}' = \mathbf{x}$) early on. From the proposition 1, *ID gradient* $(\mathbf{p}_k - \mathbf{y}_k)$ should dominate *OOD gradient* $(\mathbf{p}'_k - 1/C)$ to reliably learn ID discrimination as effective outlier synthesis relies on accurately understanding ID features. As the model gets better on ID discrimination, the overconfident nature of neural networks can often lead to high-confidence predictions for both ID and OOD (high $\mathbf{p}'_k$ and $\mathbf{p}_k$), implying $|\mathbf{p}_k - \mathbf{y}_k| < |\mathbf{p}'_k - 1/C|$. Though the nature of outlier synthesis determines $\mathbf{p}'_k$, it is desirable to avoid high-confidence predictions on challenging outliers.

Proposition 2 *The norm of a standardized feature $\tilde{\mathbf{h}} \in \mathbb{R}^m$ with $(\mu, \sigma) = (0, \sigma)$ is constrained by the upper bound $\sigma \cdot \sqrt{m-1}$.*

Proof. We begin by examining the square of the norm of standardized feature $\tilde{\mathbf{h}}$ of $\mathbf{h}$:

$$\|\tilde{\mathbf{h}}\|^2 = \sum_{u=1}^{m} \tilde{h}_u^2 = \sum_{u=1}^{m} \left(\left(\frac{h_u - \mu_h}{\sigma_h}\right) \cdot \sigma + \mu\right)^2$$

$$\|\tilde{\mathbf{h}}\|^2 = \frac{\sigma^2}{\sigma_h^2} \sum_{u=1}^{m} (h_u - \mu_h)^2 \tag{3}$$

From the definition of sample standard deviation, we have:

$$\sigma_h^2 = \frac{\sum_{u=1}^{m} (h_u - \mu_h)^2}{m-1} \Rightarrow \sigma_h^2 \cdot (m-1) = \sum_{u=1}^{m} (h_u - \mu_h)^2 \tag{4}$$

Substituting this equality into Equation 3:

$$\|\tilde{\mathbf{h}}\|^2 = \frac{\sigma^2}{\sigma_h^2} \sum_{u=1}^{m} (h_u - \mu_h)^2 = \frac{\sigma^2}{\sigma_h^2} \cdot \sigma_h^2 \cdot (m-1) = \sigma^2 \cdot (m-1)$$

$$\text{Hence, } \|\tilde{\mathbf{h}}\| = \sigma \cdot \sqrt{m-1}$$

□

We hypothesize that effective joint optimization of ID classification and outlier uncertainty requires mitigating overconfidence. The proposition 2 states that the norm of the standardized feature $\tilde{\mathbf{h}} = \mathcal{S}_h(\mathbf{h}) = \left(\left(\frac{\mathbf{h} - \mu_{\mathbf{h}}}{\sigma_{\mathbf{h}}}\right) \cdot \sigma\right)(\mu = 0)$ is constrained by the upper bound $\sigma \cdot \sqrt{m-1}$. We hypothesize that employing constrained optimization by using standardized feature $\tilde{\mathbf{h}}$ instead of raw feature $\mathbf{h}$ minimizes overconfidence. Indeed, prior works [65, 88] have shown the effectiveness of constrained optimization. Comparatively low values of $\mathbf{p}_k$ and $\mathbf{p}'_k$ can often ensure $|\mathbf{p}_k - \mathbf{y}_k| >$

$|\mathbf{p}'_k - 1/C|$. This reduction in overconfidence can assist in maintaining the appropriate balance of *ID gradient* and *OOD gradient*. Furthermore, a hyperparameter λ can be introduced to empirically achieve the desired balance in $\mathcal{L}_{\text{total}}$ such that, $\mathcal{L}_{\text{total}} = \mathcal{L}_{CE} + \lambda \cdot \mathcal{L}_{KL}$. The training-time regularization objective 1 can be expressed as:

$$\arg\min_{\theta,\gamma} \quad \underbrace{\mathcal{L}_{\text{CE}}\left(f_\theta\left(\mathcal{S}_h\left(\phi_\gamma\left(\mathbf{x}\right)\right)\right), \mathbf{y}\right)}_{\text{ID classification error}} + \lambda \cdot \underbrace{\mathcal{L}_{\text{KL}}\left(f_\theta\left(\mathcal{S}_h\left(\phi_\gamma\left(\mathbf{x}'\right)\right)\right), \mathcal{U}\right)}_{\text{Uncertainty error}} \tag{5}$$

We train our model by using this objective in Eq. (5).

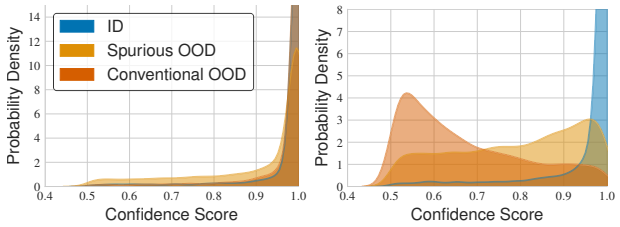
4. Experiments

OOD Datasets: The details regarding spurious OOD (of Waterbirds and CelebA) and fine-grained OOD (of Aircraft and Car) datasets are provided in Sec. A. For conventional OOD datasets under both spurious and fine-grained setup, we use NINCO [8], SUN [90], OpenImage-O [83], iNaturalist [27], and Textures [12] datasets. We use following conventional OOD datasets for CIFAR-10/100 ID datasets: MNIST [14], SVHN [58], iSUN [92] Textures [12], Places365 [108] and for ImageNet-100: SSB-Hard [81], OpenImage-O [83], iNaturalist [27], and Textures [12].

Experimental details. We adhere closely to the training procedures outlined in OpenOOD [94, 103] with a few modifications. The experiments in fine-grained settings are performed with a batch size of 32. We use ResNet-18 model in spurious and conventional settings (CIFAR-10/100), while we use ResNet-50 model in fine-grained settings. We perform experiments in the conventional setting (CIFAR-10/100) from scratch while other settings follow a fine-tuning approach. For spurious and fine-grained settings, we fine-tune the (ImageNet) pre-trained model with an initial learning rate of 0.01 for 30 epochs. For ImageNet-100 experiments, we adopt the experimental setup of DreamOOD [18]. We fine-tune the pre-trained ResNet-34 base model for 20 epochs with a batch size of 40 and a learning rate of 0.0005. For LogitNorm [88] training in the spurious setting, we set the temperature to 1. Please refer to Sec. D for complete details.

Metrics: We evaluate OOD detection using AUROC (Area Under Receiver-Operator Characteristics) and FPR@95 (False Positive Rate at 95% True Positive Rate), where higher AUROC indicates better OOD/ID discrimination and lower FPR reflects fewer ID samples misclassified as OOD.

Baselines: MSP [23], GEN [49], ODIN [45], MDS [40], MDSEns [40], TempScale [22], RMDS [67], Gram [70], EBO [48], GradNorm [31], ReAct [76], MLS [26], KLM [26], VIM [83], DICE [75], RankFeat [74], ASH [16], SHE [101], NNGuide [63], Relation [35], SCALE [91], FDBD [47], ConfBranch [15], RotPred [25], G-ODIN [29], MOS [30], VOS [17] , LogitNorm [88], CIDER [57], NPOS [79], OE [24], MixOE [102], DreamOOD [18].



(a) Baseline (b) ASCOOD

Figure 4. Visualization of confidence scores (MSP) of (a) cross-entropy baseline and (b) ASCOOD in Waterbirds benchmark. The confidence scores of ID and (spurious and conventional (iNaturalist)) OOD are relatively well-separated in case of (b) ASCOOD in comparison to (a) cross-entropy baseline.

5. Results

Fine-grained OOD detection. We assess OOD detection performance in fine-grained setting using Aircraft and Car benchmarks. Results for fine-grained OOD datasets are presented in Table 1, with performance on conventional OOD datasets deferred to Sec. G.11 and Sec. G.13. ASCOOD outperforms the nearest competitors GEN and RMDS by ∼**3** AUROC points in Aircraft datasets. Notably, many training regularization approaches (RotPred, G-ODIN, VOS, LogitNorm, CIDER, NPOS) fail to even match performance of MSP baseline in Car benchmark. On the other hand, ASCOOD achieves the best performance among all 30 competing methods, surpassing third-best rival ReAct in FPR@95 / AUROC metric by ∼**5/3** points. MixOE narrows this performance gap in Car datasets by leveraging external OOD datasets $\mathbb{D}_{\text{out}}$ (SUN datasets), though this benefit doesn't extend to Aircraft datasets under same conditions.

Spurious OOD detection. Spurious OOD detection is summarized in Table 1 using Waterbirds and CelebA benchmarks. On Waterbirds benchmark, ASCOOD outperforms all 30 competing methods by a substantial margin in both FPR@95 and AUROC metrics. Specifically, ASCOOD achieves a performance improvement over the nearest competitor Relation by ∼**59%** in FPR@95 metric. Moreover, ASCOOD sustains its superiority on the CelebA benchmark too, surpassing second-best FPR@95 metric of CIDER by ∼**15%**. Such impressive performance of ASCOOD can be partly attributed to nature of virtual outliers as they potentially contain spurious features that are present in ID samples. However, using external OOD datasets (OE) or mixing them with ID samples (MixOE) may either lack spurious features or lead to their comparatively stronger degradation. Leveraging Car datasets as external OOD datasets, both OE and MixOE outperform other training regularization methods in Waterbirds bench-

Method	Fine-grained OOD Detection		Spurious OOD Detection		Conventional OOD Detection	
	Aircraft	Car	Waterbirds	CelebA	CIFAR-100	CIFAR-10
MSP	63.79$_{\pm 5.71}$ / 80.53$_{\pm 1.61}$	58.17$_{\pm 0.99}$ / 87.12$_{\pm 0.16}$	60.41$_{\pm 1.52}$ / 77.18$_{\pm 0.66}$	56.00$_{\pm 2.73}$ / 82.53$_{\pm 1.19}$	57.49$_{\pm 0.85}$ / 78.48$_{\pm 0.42}$	29.94$_{\pm 1.32}$ / 91.06$_{\pm 0.35}$
TempScale	61.72$_{\pm 5.31}$ / 82.07$_{\pm 1.62}$	57.47$_{\pm 1.35}$ / 87.74$_{\pm 0.20}$	60.37$_{\pm 1.51}$ / 77.19$_{\pm 0.66}$	55.50$_{\pm 2.62}$ / 82.62$_{\pm 1.18}$	56.58$_{\pm 0.99}$ / 79.55$_{\pm 0.51}$	31.38$_{\pm 1.78}$ / 91.33$_{\pm 0.42}$
MDS	77.41$_{\pm 0.74}$ / 66.51$_{\pm 0.37}$	67.48$_{\pm 0.61}$ / 70.45$_{\pm 0.39}$	93.60$_{\pm 1.42}$ / 73.82$_{\pm 0.91}$	90.91$_{\pm 2.49}$ / 59.40$_{\pm 3.68}$	73.39$_{\pm 1.69}$ / 68.32$_{\pm 1.24}$	32.28$_{\pm 3.97}$ / 89.75$_{\pm 1.60}$
MDSEns	94.70$_{\pm 0.81}$ / 49.67$_{\pm 0.39}$	95.36$_{\pm 0.12}$ / 49.75$_{\pm 0.03}$	62.19$_{\pm 0.60}$ / 84.70$_{\pm 0.07}$	92.71$_{\pm 0.46}$ / 56.38$_{\pm 0.57}$	63.04$_{\pm 0.16}$ / 71.05$_{\pm 0.33}$	55.59$_{\pm 2.36}$ / 77.92$_{\pm 0.57}$
RMDS	58.40$_{\pm 4.11}$ / 86.80$_{\pm 0.44}$	51.46$_{\pm 1.20}$ / 88.25$_{\pm 0.15}$	84.06$_{\pm 4.20}$ / 71.61$_{\pm 1.44}$	88.36$_{\pm 1.89}$ / 72.04$_{\pm 1.69}$	52.95$_{\pm 0.13}$ / 82.50$_{\pm 0.10}$	24.34$_{\pm 0.74}$ / 92.55$_{\pm 0.22}$
Gram	96.01$_{\pm 0.25}$ / 43.53$_{\pm 1.42}$	92.19$_{\pm 0.08}$ / 55.74$_{\pm 0.68}$	92.71$_{\pm 0.44}$ / 66.95$_{\pm 1.87}$	72.17$_{\pm 3.63}$ / 68.46$_{\pm 2.20}$	71.55$_{\pm 1.90}$ / 61.48$_{\pm 1.03}$	77.65$_{\pm 5.39}$ / 64.28$_{\pm 2.69}$
EBO	51.22$_{\pm 3.88}$ / 85.80$_{\pm 1.25}$	59.36$_{\pm 2.64}$ / 86.83$_{\pm 0.35}$	59.17$_{\pm 1.43}$ / 77.66$_{\pm 0.79}$	55.19$_{\pm 3.10}$ / 82.50$_{\pm 1.10}$	54.76$_{\pm 1.42}$ / 80.84$_{\pm 0.60}$	38.82$_{\pm 4.25}$ / 91.63$_{\pm 0.76}$
GradNorm	90.49$_{\pm 1.67}$ / 70.78$_{\pm 1.37}$	86.86$_{\pm 0.70}$ / 72.13$_{\pm 0.37}$	72.16$_{\pm 2.79}$ / 79.79$_{\pm 1.47}$	62.57$_{\pm 3.88}$ / 77.16$_{\pm 2.26}$	84.21$_{\pm 2.34}$ / 69.60$_{\pm 1.23}$	91.07$_{\pm 1.76}$ / 59.41$_{\pm 2.67}$
ReAct	60.11$_{\pm 6.33}$ / 83.62$_{\pm 1.22}$	45.46$_{\pm 2.65}$ / 88.89$_{\pm 0.29}$	56.37$_{\pm 1.95}$ / 78.80$_{\pm 0.89}$	55.14$_{\pm 2.85}$ / 82.77$_{\pm 0.92}$	52.25$_{\pm 1.48}$ / 81.46$_{\pm 0.55}$	41.09$_{\pm 6.33}$ / 91.06$_{\pm 1.04}$
MLS	52.00$_{\pm 3.80}$ / 85.99$_{\pm 1.25}$	59.19$_{\pm 2.48}$ / 87.42$_{\pm 0.35}$	59.17$_{\pm 1.43}$ / 77.69$_{\pm 0.68}$	55.19$_{\pm 3.10}$ / 82.52$_{\pm 1.11}$	54.92$_{\pm 1.35}$ / 80.65$_{\pm 0.57}$	38.79$_{\pm 4.19}$ / 91.52$_{\pm 0.73}$
KLM	82.29$_{\pm 4.20}$ / 80.74$_{\pm 2.01}$	66.45$_{\pm 1.43}$ / 85.50$_{\pm 0.15}$	97.68$_{\pm 0.09}$ / 46.97$_{\pm 0.47}$	98.61$_{\pm 0.16}$ / 53.12$_{\pm 0.97}$	74.15$_{\pm 1.15}$ / 76.46$_{\pm 0.50}$	18.16$_{\pm 2.10}$ / 94.69$_{\pm 0.87}$
VIM	57.51$_{\pm 4.10}$ / 79.21$_{\pm 0.94}$	54.80$_{\pm 1.23}$ / 82.03$_{\pm 0.16}$	39.66$_{\pm 1.21}$ / 85.32$_{\pm 0.63}$	58.25$_{\pm 1.67}$ / 82.73$_{\pm 0.29}$	49.92$_{\pm 1.06}$ / 81.81$_{\pm 0.76}$	24.19$_{\pm 0.31}$ / 93.64$_{\pm 0.16}$
DICE	70.79$_{\pm 4.24}$ / 72.20$_{\pm 3.31}$	72.27$_{\pm 0.87}$ / 77.45$_{\pm 0.31}$	56.40$_{\pm 2.22}$ / 84.91$_{\pm 1.55}$	50.52$_{\pm 2.50}$ / 80.90$_{\pm 2.15}$	54.55$_{\pm 0.45}$ / 80.93$_{\pm 0.30}$	53.31$_{\pm 5.29}$ / 83.70$_{\pm 2.16}$
RankFeat	62.00$_{\pm 4.57}$ / 82.13$_{\pm 1.74}$	87.45$_{\pm 3.73}$ / 62.05$_{\pm 3.57}$	70.47$_{\pm 8.53}$ / 67.50$_{\pm 6.04}$	74.48$_{\pm 2.51}$ / 67.38$_{\pm 3.39}$	67.90$_{\pm 0.96}$ / 68.52$_{\pm 1.32}$	54.44$_{\pm 9.49}$ / 77.48$_{\pm 5.48}$
ASH	86.41$_{\pm 3.76}$ / 73.98$_{\pm 3.29}$	80.25$_{\pm 4.34}$ / 74.73$_{\pm 3.93}$	36.69$_{\pm 1.16}$ / 87.03$_{\pm 0.62}$	58.49$_{\pm 4.15}$ / 80.98$_{\pm 0.64}$	52.84$_{\pm 1.20}$ / 82.26$_{\pm 0.54}$	70.55$_{\pm 5.50}$ / 85.27$_{\pm 2.05}$
SHE	78.90$_{\pm 2.46}$ / 76.13$_{\pm 1.45}$	86.78$_{\pm 1.34}$ / 71.16$_{\pm 0.64}$	66.03$_{\pm 2.08}$ / 82.99$_{\pm 1.82}$	55.19$_{\pm 4.01}$ / 78.55$_{\pm 2.02}$	62.08$_{\pm 2.73}$ / 77.99$_{\pm 1.09}$	63.71$_{\pm 5.22}$ / 86.18$_{\pm 1.20}$
GEN	50.81$_{\pm 5.85}$ / 86.41$_{\pm 1.36}$	56.98$_{\pm 2.40}$ / 88.05$_{\pm 0.51}$	60.37$_{\pm 1.51}$ / 77.19$_{\pm 0.66}$	55.50$_{\pm 2.62}$ / 82.62$_{\pm 1.18}$	55.11$_{\pm 1.65}$ / 80.48$_{\pm 0.93}$	32.49$_{\pm 1.17}$ / 91.72$_{\pm 0.57}$
NNGuide	52.23$_{\pm 3.23}$ / 85.13$_{\pm 1.42}$	61.12$_{\pm 2.45}$ / 85.86$_{\pm 0.34}$	53.69$_{\pm 1.40}$ / 80.39$_{\pm 0.75}$	52.32$_{\pm 2.39}$ / 83.22$_{\pm 1.02}$	51.89$_{\pm 1.35}$ / 82.33$_{\pm 0.66}$	39.64$_{\pm 3.91}$ / 91.39$_{\pm 0.57}$
Relation	61.35$_{\pm 5.78}$ / 83.05$_{\pm 1.90}$	56.25$_{\pm 1.30}$ / 86.77$_{\pm 0.84}$	31.71$_{\pm 0.68}$ / 88.35$_{\pm 0.24}$	62.19$_{\pm 2.65}$ / 81.78$_{\pm 0.45}$	53.84$_{\pm 0.32}$ / 81.60$_{\pm 0.34}$	26.59$_{\pm 0.64}$ / 92.49$_{\pm 0.22}$
SCALE	90.49$_{\pm 1.67}$ / 70.78$_{\pm 1.37}$	79.78$_{\pm 1.19}$ / 73.78$_{\pm 0.35}$	36.49$_{\pm 1.94}$ / 91.89$_{\pm 0.66}$	75.82$_{\pm 3.25}$ / 72.58$_{\pm 1.52}$	53.20$_{\pm 1.27}$ / 81.97$_{\pm 0.54}$	63.15$_{\pm 5.95}$ / 87.38$_{\pm 1.51}$
FDBD	58.46$_{\pm 6.07}$ / 83.71$_{\pm 1.50}$	50.30$_{\pm 2.59}$ / 88.18$_{\pm 0.26}$	50.07$_{\pm 1.19}$ / 80.34$_{\pm 0.57}$	56.07$_{\pm 3.26}$ / 83.16$_{\pm 0.50}$	51.66$_{\pm 0.41}$ / 81.23$_{\pm 0.41}$	23.00$_{\pm 0.80}$ / 93.44$_{\pm 0.26}$
ConfBranch	91.24$_{\pm 0.97}$ / 41.85$_{\pm 0.70}$	80.63$_{\pm 1.80}$ / 62.19$_{\pm 0.76}$	52.09$_{\pm 5.40}$ / 85.58$_{\pm 2.42}$	57.02$_{\pm 1.53}$ / 80.39$_{\pm 0.52}$	77.97$_{\pm 1.16}$ / 63.82$_{\pm 2.02}$	21.40$_{\pm 1.07}$ / 93.30$_{\pm 0.50}$
RotPred	62.04$_{\pm 1.30}$ / 81.98$_{\pm 0.46}$	61.86$_{\pm 2.18}$ / 85.00$_{\pm 0.23}$	42.06$_{\pm 1.55}$ / 85.32$_{\pm 0.94}$	52.74$_{\pm 2.52}$ / 83.26$_{\pm 0.47}$	35.57$_{\pm 1.04}$ / 88.09$_{\pm 0.33}$	12.80$_{\pm 0.69}$ / 96.36$_{\pm 0.12}$
G-ODIN	58.43$_{\pm 4.04}$ / 83.09$_{\pm 0.06}$	64.34$_{\pm 6.38}$ / 85.94$_{\pm 1.08}$	58.65$_{\pm 17.87}$ / 80.87$_{\pm 4.18}$	78.17$_{\pm 2.96}$ / 58.68$_{\pm 1.56}$	37.03$_{\pm 0.43}$ / 88.49$_{\pm 0.36}$	20.02$_{\pm 0.91}$ / 95.79$_{\pm 0.19}$
VOS	57.31$_{\pm 1.34}$ / 85.37$_{\pm 0.27}$	63.16$_{\pm 2.77}$ / 86.42$_{\pm 0.42}$	57.66$_{\pm 3.13}$ / 78.65$_{\pm 1.26}$	54.67$_{\pm 2.02}$ / 82.56$_{\pm 0.65}$	53.28$_{\pm 4.01}$ / 81.42$_{\pm 1.95}$	38.08$_{\pm 3.29}$ / 92.04$_{\pm 0.52}$
LogitNorm	70.32$_{\pm 5.73}$ / 79.16$_{\pm 1.67}$	62.48$_{\pm 1.90}$ / 84.92$_{\pm 0.49}$	64.18$_{\pm 4.19}$ / 73.87$_{\pm 0.45}$	100.00$_{\pm 0.00}$ / 81.61$_{\pm 1.99}$	52.42$_{\pm 1.67}$ / 80.70$_{\pm 1.15}$	13.98$_{\pm 1.33}$ / 96.54$_{\pm 0.45}$
CIDER	88.99$_{\pm 1.88}$ / 54.08$_{\pm 1.50}$	88.95$_{\pm 1.07}$ / 54.30$_{\pm 1.31}$	49.34$_{\pm 17.95}$ / 84.61$_{\pm 7.94}$	50.38$_{\pm 4.98}$ / 85.35$_{\pm 0.92}$	61.67$_{\pm 1.69}$ / 77.22$_{\pm 0.89}$	20.23$_{\pm 2.90}$ / 94.49$_{\pm 1.18}$
NPOS	68.71$_{\pm 10.57}$ / 72.02$_{\pm 6.07}$	89.09$_{\pm 0.53}$ / 55.95$_{\pm 0.64}$	42.88$_{\pm 5.95}$ / 89.01$_{\pm 1.83}$	54.35$_{\pm 5.25}$ / 82.06$_{\pm 4.26}$	52.09$_{\pm 2.02}$ / 84.06$_{\pm 0.99}$	22.65$_{\pm 1.10}$ / 93.51$_{\pm 0.18}$
OE	56.82$_{\pm 2.01}$ / 82.86$_{\pm 0.48}$	60.30$_{\pm 3.04}$ / 85.60$_{\pm 1.73}$	33.26$_{\pm 6.00}$ / 92.63$_{\pm 1.14}$	100.0$_{\pm 0.00}$ / 70.24$_{\pm 3.83}$	46.82$_{\pm 4.65}$ / 86.84$_{\pm 1.44}$	19.50$_{\pm 3.72}$ / 92.82$_{\pm 1.80}$
MixOE	70.61$_{\pm 12.18}$ / 83.47$_{\pm 2.41}$	42.04$_{\pm 0.74}$ / 90.19$_{\pm 0.23}$	42.13$_{\pm 4.26}$ / 88.09$_{\pm 3.18}$	77.52$_{\pm 2.05}$ / 74.31$_{\pm 0.79}$	68.76$_{\pm 3.98}$ / 74.86$_{\pm 2.42}$	52.09$_{\pm 6.20}$ / 90.09$_{\pm 0.64}$
ASCOOD	**47.94$_{\pm 5.38}$ / 89.75$_{\pm 1.01}$**	**40.76$_{\pm 1.13}$ / 91.86$_{\pm 0.20}$**	**11.37$_{\pm 0.42}$ / 97.48$_{\pm 0.10}$**	**42.62$_{\pm 0.76}$ / 86.50$_{\pm 0.89}$**	**29.90$_{\pm 0.76}$ / 91.35$_{\pm 0.13}$**	**7.69$_{\pm 0.29}$ / 98.32$_{\pm 0.07}$**

Table 1. OOD detection results (FPR@95↓ / AUROC↑) on fine-grained, spurious, and conventional OOD detection. Best results are formatted as **bold** and second-best results are formatted as underline. Same formatting is applied to subsequent tables. See Appendix G.

Conventional setting	Spurious setting	Fine-grained setting
CIFAR-10/100 [37, 38]	CelebA [51]	Aircraft (ID/OOD categories = 90/10) [55]
ImageNet-100 [13]	Waterbirds [69]	Car (ID/OOD categories = 150/46) [36]

Table 2. ID datasets used in our experiments (See A).

mark, yet still fall short of ASCOOD's performance. Quantitatively, ASCOOD demonstrates a substantial advantage, outperforming OE and MixOE by **66%** and **73%** in the FPR@95 metric on the Waterbirds datasets. This superior performance is consistent across datasets, with ASCOOD outperforming OE by **57%** and MixOE by **45%** on CelebA datasets. Figures 4a and 4b visualize confidence scores for (a) cross-entropy baseline and (b) ASCOOD in Waterbirds benchmark. These plots demonstrate that ASCOOD achieves relatively more pronounced separation of confidence scores between ID and spurious OOD, as well as between ID and conventional OOD (iNaturalist).

Conventional OOD detection. We evaluate OOD detection performance in conventional setting using the CIFAR-10/100 benchmarks as presented in Table 1. Consistent with results in the fine-grained and spurious setting, ASCOOD outperforms all 30 competing methods by a significant margin. Specifically, ASCOOD exceeds the performance of the strong RotPred baseline on the CIFAR-100 benchmark, improving the FPR@95 metric by ∼**16%**. Additionally, an AUROC improvement of ∼**3** points is observed between ASCOOD and RotPred. While CIFAR-10 is considered a comparatively easier benchmark on which many prior methods perform well, ASCOOD still demonstrates superior performance across both metrics. Apart from CIFAR benchmarks, we conduct experiments in large-scaling settings using ImageNet-100 ID datasets following the experimental settings of Dream-OOD [18]. The results presented in Table 3 demonstrate the strong empirical effectiveness of ASCOOD in large-scale settings. Furthermore, when evaluated on SSB-Hard [81], ASCOOD achieves the highest AUROC score (**83.91**), surpassing the second-best score of DreamOOD (83.30). (See Sec. G.5 and Sec. G.6 for complete results.)

ODIN vs. i-ODIN. Unlike ODIN, which perturbs all color channel intensities, we propose i-ODIN that modifies the variable number of significant color channel intensities (determined via pixel attribution) of the image. We compare ODIN and i-ODIN in challenging cases as shown in Table 4 which demonstrate significant gain of i-ODIN over ODIN

establishing superiority of the former. For instance, i-ODIN outperforms ODIN in FPR@95 metric by **20%** in CIFAR-10 (ID) vs CIFAR-100 (OOD) and by **33%** in CIFAR-100 (ID) vs TIN (OOD). Furthermore, a non-trivial performance improvement of i-ODIN over ODIN is also observed in fine-grained setting, especially in the FPR@95 metric. Specifically, we find perturbing only the single most significant color channel intensity yields substantially better performance than perturbing all color channel intensities. However, in trivial cases involving only two classes (e.g., Waterbirds and CelebA datasets), this improvement is not observed. Please see Section G.1 for complete results.

Accuracy: It is undesirable to trade off accuracy with OOD detection performance. ASCOOD achieves accuracies of ∼87.27%, 76.63%, 94.95%, 93.59%, and 96.58% on ImageNet-100, CIFAR-100, CIFAR-10, Waterbirds, and CelebA respectively, closely aligning with the baseline accuracies of ∼87.33%, 77.25%, 95.06%, 93.72%, and 96.72%. In fine-grained setting, ASCOOD achieves slightly better accuracies of ∼**94.20**% and **89.61**% in Car and Aircraft datasets respectively bettering baseline accuracies of ∼92.73% and 87.85%. It shows efficacy of ASCOOD in enhancing OOD detection without harming accuracy.

5.1. Ablation studies

Outlier synthesis. For a sufficiently high value of α used in synthesizing virtual outliers $\mathbf{x}'$, the invariant features of $\mathbf{x}$ are destroyed, irrespective of whether gradient addition or subtraction is employed. As the subtraction of the gradient reduces the logit value associated with the true class in resulting outliers, the model tends to exhibit increased uncertainty towards them. Consequently, the resulting outliers are less challenging and offer limited potential for enhancing predictive uncertainty towards OOD samples. Therefore, substantial performance gains are not anticipated with this approach. Conversely, gradient addition increases the true class logit while simultaneously compromising invariant features (with high α), creating an opportunity to improve predictive uncertainty towards outliers. By incentivizing predictive uncertainty for these virtual outliers, the model learns an improved discrimination between known and unknown data. We empirically analyze impact of these outlier synthesis strategies along with invariant pixel shuffling on fine-grained OOD detection, reporting results on Aircraft and Car datasets in Table 5. The results verify that gradient addition is a superior choice for outlier synthesis.

Method	iNaturalist	Textures	OpenImage	Average
MSP	21.09 / 94.87	62.82 / 83.45	39.82 / 88.08	41.24 / 88.80
ODIN	21.02 / 95.43	79.42 / 80.92	50.42 / 85.60	50.29 / 87.32
EBO	22.24 / 94.03	74.67 / 81.04	43.98 / 86.66	46.96 / 87.24
ReAct	20.89 / 94.68	65.98 / 82.42	41.27 / 87.41	42.71 / 88.17
SHE	29.51 / 93.28	82.47 / 82.13	61.76 / 81.77	57.91 / 85.72
GEN	18.64 / 94.62	65.60 / 82.48	39.51 / 88.45	41.25 / 88.52
NNGuide	19.73 / 95.59	71.31 / 84.52	43.00 / 87.52	44.68 / 89.21
SCALE	**12.91 / 97.32**	54.13 / 89.77	38.58 / 89.96	35.21 / 92.35
RotPred	19.64 / 93.57	50.51 / 88.69	37.44 / 88.64	35.87 / 90.30
VOS	21.20 / 94.25	64.96 / 83.72	36.82 / 88.26	40.99 / 88.74
LogitNorm	18.38 / 95.75	49.96 / 87.23	34.69 / 89.51	34.34 / 90.83
CIDER	24.91 / 95.03	**21.84 / 96.20**	46.29 / 89.41	31.01 / 93.54
Dream-OOD (EBO)	14.47 / 96.09	60.73 / 84.79	32.67 / 90.16	35.96 / 90.35
ASCOOD (EBO)	18.11 / 95.73	25.20 / 94.40	**26.04 / 91.95**	**23.12 / 94.02**

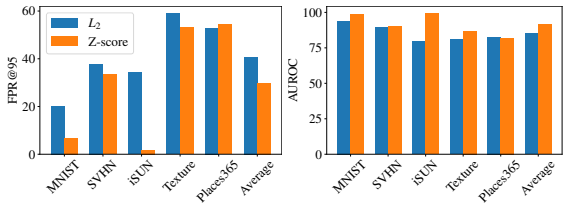
Table 3. Conventional OOD detection (FPR@95 ↓/ AUROC ↑) in large-scale setting using ImageNet-100 datasets.

Benchmark	Method	CIFAR-100		TIN	
		FPR@95 ↓	AUROC ↑	FPR@95 ↓	AUROC ↑
CIFAR-10	ODIN	77.00$_{\pm 5.74}$	82.18$_{\pm 1.87}$	75.38$_{\pm 6.42}$	83.55$_{\pm 1.84}$
	i-ODIN	**61.33**$_{\pm 1.18}$	**86.87**$_{\pm 0.14}$	**50.20**$_{\pm 2.39}$	**88.96**$_{\pm 0.25}$
Benchmark	**Method**	**Car**		**Aircraft**	
		FPR@95 ↓	AUROC ↑	FPR@95 ↓	AUROC ↑
Fine-grained setting	ODIN	64.78$_{\pm 1.28}$	86.41$_{\pm 0.29}$	54.55$_{\pm 5.54}$	**86.23**$_{\pm 1.47}$
	i-ODIN	**58.21**$_{\pm 2.41}$	**87.80**$_{\pm 0.36}$	**51.81**$_{\pm 4.82}$	**86.21**$_{\pm 1.28}$

Table 4. ODIN vs. i-ODIN in challenging cases.

$\mathbf{x}'$	Car		Aircraft	
	FPR@95 ↓	AUROC ↑	FPR@95 ↓	AUROC ↑
$\mathbf{x}' = \psi(\texttt{shuffle}(\mathbf{x}_{\text{inv}}), \mathbf{e})$	63.84$_{\pm 3.21}$	85.38$_{\pm 0.29}$	47.98$_{\pm 2.04}$	83.87$_{\pm 0.38}$
$\mathbf{x}' = \mathbf{x} - \alpha \cdot \mathbf{G}_{\text{inv}}$	60.20$_{\pm 1.91}$	86.27$_{\pm 0.34}$	50.15$_{\pm 4.80}$	83.64$_{\pm 0.82}$
$\mathbf{x}' = \mathbf{x} + \alpha \cdot \mathbf{G}_{\text{inv}}$	**40.76**$_{\pm 1.13}$	**91.86**$_{\pm 0.20}$	**47.94**$_{\pm 5.38}$	**89.75**$_{\pm 1.01}$

Table 5. Ablation of outlier synthesis methods on challenging fine-grained OOD detection. (See Sec. D.5 for complete results.)



(a) FPR@95 in various datasets. (b) AUROC in various datasets.

Figure 5. Comparison of L_2 normalization and Z-score normalization in terms of FPR@95 and AUROC in CIFAR-100 datasets.

Standardized feature space: As L_2 normalization techniques [62, 65, 88] have been explored in prior works, its empirical effectiveness in comparison to z-score normalization is unknown. We make this comparison in CIFAR-100 benchmark with bar charts in Figure 5a and Figure 5b using invariant pixel shuffling for outlier synthesis. We observe that L_2 normalization yields average OOD detection performance (FPR@95/AUROC: $40.81_{\pm 3.06}$/$85.26_{\pm 3.10}$) which is inferior to the average performance of ASCOOD ($29.90_{\pm 0.76}$/$91.35_{\pm 0.13}$). Furthermore, we also conduct experiments in fine-grained settings using gradient addition for outlier synthesis. The results, presented in Table 6, consistently demonstrate z-score normalization to be superior. For instance, z-score normalization leads to improvement of **19**% and **31**% in FPR@95 metric in Aircraft and Car datasets, respectively. Further-

$\mathbb{D}_{\text{in}}$	$\mathcal{L}_{\text{OOD}}$	Feature space	**Fine-grained OOD**	**Conventional OOD**
Aircraft	$\mathcal{L}_{\text{energy}}$	$\mathbf{h}$	$72.57_{\pm 2.52}$ / $78.13_{\pm 0.81}$	$3.21_{\pm 0.04}$ / $99.09_{\pm 0.04}$
		$\mathbf{h} / \|\|\mathbf{h}\|\|_2$	$57.56_{\pm 8.62}$ / $85.17_{\pm 1.10}$	$40.50_{\pm 7.75}$ / $88.10_{\pm 1.77}$
		$(\mathbf{h}-\mu_{\mathbf{h}})\cdot\sigma/\sigma_{\mathbf{h}}$	$58.70_{\pm 4.90}$ / $84.47_{\pm 0.71}$	$84.63_{\pm 8.10}$ / $53.51_{\pm 6.27}$
	$\mathcal{L}_{\text{KL}}$	$\mathbf{h}$	$55.21_{\pm 5.98}$ / $\underline{88.73}_{\pm 0.89}$	$6.84_{\pm 0.95}$ / $98.47_{\pm 0.11}$
		$\mathbf{h} / \|\|\mathbf{h}\|\|_2$	$\underline{54.30}_{\pm 5.87}$ / $85.70_{\pm 1.48}$	$\underline{2.58}_{\pm 0.06}$ / $\underline{99.37}_{\pm 0.04}$
		$(\mathbf{h}-\mu_{\mathbf{h}})\cdot\sigma/\sigma_{\mathbf{h}}$	$\mathbf{47.94}_{\pm 5.38}$ / $\mathbf{89.75}_{\pm 1.01}$	$\mathbf{0.55}_{\pm 0.07}$ / $\mathbf{99.84}_{\pm 0.02}$
Car	$\mathcal{L}_{\text{energy}}$	$\mathbf{h}$	$54.15_{\pm 1.23}$ / $88.09_{\pm 0.12}$	$6.98_{\pm 0.55}$ / $98.73_{\pm 0.08}$
		$\mathbf{h} / \|\|\mathbf{h}\|\|_2$	$61.80_{\pm 0.81}$ / $86.57_{\pm 0.04}$	$7.16_{\pm 2.87}$ / $98.70_{\pm 0.39}$
		$(\mathbf{h}-\mu_{\mathbf{h}})\cdot\sigma/\sigma_{\mathbf{h}}$	$58.84_{\pm 1.05}$ / $86.46_{\pm 0.37}$	$\mathbf{3.07}_{\pm 1.02}$ / $\mathbf{99.41}_{\pm 0.18}$
	$\mathcal{L}_{\text{KL}}$	$\mathbf{h}$	$54.89_{\pm 2.51}$ / $\underline{90.21}_{\pm 0.21}$	$14.32_{\pm 2.43}$ / $96.58_{\pm 0.48}$
		$\mathbf{h} / \|\|\mathbf{h}\|\|_2$	$\underline{51.86}_{\pm 4.10}$ / $89.70_{\pm 0.23}$	$9.55_{\pm 2.82}$ / $97.82_{\pm 0.39}$
		$(\mathbf{h}-\mu_{\mathbf{h}})\cdot\sigma/\sigma_{\mathbf{h}}$	$\mathbf{40.76}_{\pm 1.13}$ / $\mathbf{91.86}_{\pm 0.20}$	$\underline{4.28}_{\pm 0.53}$ / $\underline{98.79}_{\pm 0.12}$

Table 6. Ablation study of feature space $\mathbf{h}$ and uncertainty loss $\mathcal{L}_{\text{OOD}}$ in fine-grained setting in FPR@95↓ / AUROC↑.

more, we can also observe that employing $\mathcal{L}_{\text{KL}}$ for $\mathcal{L}_{\text{OOD}}$ yields superior results in comparison to $\mathcal{L}_{\text{energy}}$ (used in prior works [17, 18]) in fine-grained setting. Please refer Sec. D for additional empirical studies.

6. Related Works

OOD detection Since Nguyen *et al*. [59] highlighted the overconfidence of neural networks towards OOD data, numerous studies have proposed to mitigate this issue. Post-hoc methods for OOD detection offer a practical solution by operating on pre-trained models. Hendrycks *et al*. [23] uses the maximum softmax probability (MSP) as a measure of confidence. Liang *et al*. [45] improves upon MSP by incorporating temperature scaling and input preprocessing. Liu *et al*. [48] utilizes energy function derived from the softmax denominator. Lee *et al*. [40] estimates class-conditional Gaussian distributions in feature space and uses maximum Mahalanobis distance to all class centroids as a measure of OOD uncertainty. Methods offered by Sun *et al*. [76] and Ahn *et al*. [1] analyze the activation patterns of neurons to identify and leverage distinctive features for OOD detection. Yang *et al*. [94] and Zhang *et al*. [103] provide the comprehensive benchmark of such many other post-hoc methods [26, 67, 77, 101].

Beyond post-hoc methods, various regularization strategies have been explored for OOD detection. DeVries *et al*. [15] proposes learning confidence estimates by attaching an auxiliary branch to a pre-trained classifier. Hsu *et al*. [29] proposes novel confidence scoring decomposition and input preprocessing. Huang *et al*. [30] describes a group-based framework to refine decision boundaries. Yu *et al*. [100] demonstrates the importance of feature norms, while Wei *et al*. [88] and Regmi *et al*. [65] study the utility of normalization in OOD Detection. Recent works also investigate the role of contrastive learning [53, 57, 66, 71, 78] and self-supervised learning [25] for OOD detection. More recent advancements include leveraging masked image modeling [42], balancing energy regularization [11], decoupling MaxLogit [106], exploring binary neuron patterns [61], and applying uncertainty-aware optimal transport [52]. Additionally, there has been growing interest in simultaneously addressing OOD detection and generalization [5, 50, 95].

Several works such as OE [24], MCD [99], and UDG [93] tackle OOD detection by incorporating external OOD data during training. Though similar works [64, 80, 102] do the same to enhance OOD detection in fine-grained settings, curating OOD datasets that don't overlap with ID and ensuring their diversity [32, 98, 109] can be a significant hurdle. To avoid reliance on real outlier data, few recent works [17, 21, 41, 79] synthesize virtual outliers in feature space but such approach is computationally expensive. Recent studies have increasingly explored foundation models for OOD detection [6, 20, 43, 84, 105]. In contrast, our approach synthesizes virtual outliers in image space from ID data without relying on foundation models, akin to VoSo [60]. While Roy *et al*. [68] and Ahmed *et al*. [68] address challenging scenarios, their studies are limited to cases where ID and OOD share semantic similarities. There is a scarcity of studies addressing spurious correlations in the context of OOD detection. Ming *et al*. [56], Zhang *et al*. [104], Kirby [34] and BackMix [87] share some similarities with our work in terms of motivation and goal. Ming *et al*. [56] first analyze the impact of spurious settings in of OOD detection. Zhang *et al*. [104] further extend this analysis and propose a reweighting solution. In contrast to these works, our approach incorporates virtual outlier synthesis using pixel attribution. Similar to Kirby [34] and BackMix [87], we synthesize virtual outliers by removing invariant features. But, our work shows gradient addition is superior; it not only destroys these features but also makes resulting outliers challenging. More broadly, our work aims to address spurious, fine-grained, as well as conventional OOD inputs comprehensively within a unified framework.

7. Conclusion

In this work, we introduce a novel training method **ASCOOD** designed to improve OOD detection in both conventional and challenging cases. ASCOOD trains the model by incentivizing joint optimization of ID classification and predictive uncertainty towards virtual outliers. ASCOOD synthesizes virtual outliers by approximately destroying invariant features from ID images. These invariant features are determined by the pixel attribution method using the model being learned. For effective dual optimization, it employs constrained optimization in a standardized feature space. By mitigating the impact of spurious correlations and promoting the capture of fine-grained attributes, ASCOOD demonstrates improved performance in spurious, fine-grained, and conventional setups, as evidenced by extensive experiments across seven datasets. Importantly, ASCOOD operates without relying on external OOD datasets, making it a promising approach for OOD detection.

References

[1] Yong Hyun Ahn, Gyeong-Moon Park, and Seong Tae Kim. Line: Out-of-distribution detection by leveraging important neurons. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023. 8

[2] Amirhossein Ansari, Ke Wang, and Pulei Xiong. Negrefine: Refining negative label-based zero-shot ood detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2025. 2

[3] David Baehrens, Timon Schroeter, Stefan Harmeling, Motoaki Kawanabe, Katja Hansen, and Klaus-Robert Müller. How to explain individual classification decisions. *The Journal of Machine Learning Research (JMLR)*, 2010. 3

[4] Haoyue Bai, Gregory Canal, Xuefeng Du, Jeongyeol Kwon, Robert D Nowak, and Yixuan Li. Feed two birds with one scone: Exploiting wild data for both out-of-distribution generalization and detection. In *International Conference on Machine Learning (ICML)*, 2023. 2

[5] Haoyue Bai, Jifan Zhang, and Robert D Nowak. AHA: Human-assisted out-of-distribution generalization and detection. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2024. 8

[6] Yichen Bai, Zongbo Han, Bing Cao, Xiaoheng Jiang, Qinghua Hu, and Changqing Zhang. Id-like prompt learning for few-shot out-of-distribution detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024. 8

[7] Sara Beery, Grant Van Horn, and Pietro Perona. Recognition in terra incognita. In *European Conference on Computer Vision (ECCV)*, 2018. 2

[8] Julian Bitterwolf, Maximilian Müller, and Matthias Hein. In or out? fixing imagenet out-of-distribution detection evaluation. *arXiv preprint arXiv:2306.00826*, 2023. 5

[9] Jiankang Chen, Ling Deng, Zhiyong Gan, Wei-Shi Zheng, and Ruixuan Wang. Fodfom: Fake outlier data by foundation models creates stronger visual out-of-distribution detector. In *ACM Multimedia 2024*, 2024. 2

[10] Yuyan Chen, Nico Lang, B. Christian Schmidt, Aditya Jain, Yves Basset, Sara Beery, Maxim Larrivée, and David Rolnick. Open-insect: Benchmarking open-set recognition of novel species in biodiversity monitoring. In *Advances in Neural Information Processing Systems (NeurIPS), Datasets and Benchmarks Track*, 2025. 1

[11] Hyunjun Choi, Hawook Jeong, and Jin Young Choi. Balanced energy regularization loss for out-of-distribution detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023. 8

[12] Mircea Cimpoi, Subhransu Maji, Iasonas Kokkinos, Sammy Mohamed, and Andrea Vedaldi. Describing textures in the wild. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2014. 5

[13] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. ImageNet: A large-scale hierarchical image database. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*, 2009. 6

[14] Li Deng. The mnist database of handwritten digit images for machine learning research. *IEEE Signal Processing Magazine*, 2012. 5

[15] Terrance DeVries and Graham W Taylor. Learning confidence for out-of-distribution detection in neural networks. *arXiv preprint arXiv:1802.04865*, 2018. 5, 8

[16] Andrija Djurisic, Nebojsa Bozanic, Arjun Ashok, and Rosanne Liu. Extremely simple activation shaping for out-of-distribution detection. In *International Conference on Learning Representations (ICLR)*, 2023. 5

[17] Xuefeng Du, Zhaoning Wang, Mu Cai, and Yixuan Li. Vos: Learning what you don't know by virtual outlier synthesis. In *International Conference on Learning Representations (ICLR)*, 2022. 5, 8

[18] Xuefeng Du, Yiyou Sun, Jerry Zhu, and Yixuan Li. Dream the impossible: Outlier imagination with diffusion models. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2023. 2, 5, 6, 8

[19] Kun Fang, Qinghua Tao, Kexin Lv, Mingzhen He, Xiaolin Huang, and Jie Yang. Kernel pca for out-of-distribution detection. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2024. 2

[20] Ruiyuan Gao, Chenchen Zhao, Lanqing Hong, and Qiang Xu. Diffguard: Semantic mismatch-guided out-of-distribution detection using pre-trained diffusion models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2023. 8

[21] Mingrong Gong, Chaoqi Chen, Qingqiang Sun, Yue Wang, and Hui Huang. Out-of-distribution detection with prototypical outlier proxy. In *Proceedings of the AAAI conference on artificial intelligence (AAAI)*, 2024. 8

[22] Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q Weinberger. On calibration of modern neural networks. In *International Conference on Machine Learning (ICML)*, 2017. 5

[23] Dan Hendrycks and Kevin Gimpel. A baseline for detecting misclassified and out-of-distribution examples in neural networks. In *International Conference on Learning Representations (ICLR)*, 2017. 5, 8

[24] Dan Hendrycks, Mantas Mazeika, and Thomas Dietterich. Deep anomaly detection with outlier exposure. In *International Conference on Learning Representations (ICLR)*, 2019. 5, 8

[25] Dan Hendrycks, Mantas Mazeika, Saurav Kadavath, and Dawn Song. Using self-supervised learning can improve model robustness and uncertainty. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2019. 5, 8

[26] Dan Hendrycks, Steven Basart, Mantas Mazeika, Mohammadreza Mostajabi, Jacob Steinhardt, and Dawn Song. Scaling out-of-distribution detection for real-world settings. In *International Conference on Machine Learning (ICML)*, 2022. 5, 8

[27] 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 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018. 5

[28] Yen-Chi Hsu, Cheng-Yao Hong, Ding-Jie Chen, Ming-Sui Lee, Davi Geiger, and Tyng-Luh Liu. Fine-grained visual recognition with batch confusion norm. *arXiv preprint arXiv:1910.12423*, 2019. 1

[29] Yen-Chang Hsu, Yilin Shen, Hongxia Jin, and Zsolt Kira. Generalized odin: Detecting out-of-distribution image without learning from out-of-distribution data. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020. 5, 8

[30] Rui Huang and Yixuan Li. Mos: Towards scaling out-of-distribution detection for large semantic space. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021. 5, 8

[31] Rui Huang, Andrew Geng, and Yixuan Li. On the importance of gradients for detecting distributional shifts in the wild. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2021. 5

[32] Wenyu Jiang, Hao Cheng, MingCai Chen, Chongjun Wang, and Hongxin Wei. DOS: Diverse outlier sampling for out-of-distribution detection. In *The Twelfth International Conference on Learning Representations (ICLR)*, 2024. 2, 8

[33] Naveen Karunanayake, Suranga Seneviratne, and Sanjay Chawla. Craft: Class ranking aware fine-tuning for enhanced out-of-distribution detection. In *Proceedings of the Winter Conference on Applications of Computer Vision (WACV)*, 2025. 2

[34] Jaeyoung Kim, Seo Taek Kong, Dongbin Na, and Kyu-Hwan Jung. Key feature replacement of in-distribution samples for out-of-distribution detection. In *Proceedings of the AAAI conference on artificial intelligence (AAAI)*, 2023. 2, 8

[35] Jang-Hyun Kim, Sangdoo Yun, and Hyun Oh Song. Neural relation graph: a unified framework for identifying label noise and outlier data. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2024. 5

[36] Jonathan Krause, Michael Stark, Jia Deng, and Li Fei-Fei. 3d object representations for fine-grained categorization. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV) Workshops*, 2013. 6

[37] Alex Krizhevsky, Geoffrey Hinton, et al. Learning multiple layers of features from tiny images. 2009. 6

[38] Alex Krizhevsky, Vinod Nair, and Geoffrey Hinton. CIFAR-10 and CIFAR-100 datasets. 2009. 6

[39] Gitaek Kwon, Jaeyoung Kim, Hong-Jun Choi, Byung-Moo Yoon, Sungchul Choi, and Kyu-Hwan Jung. Improving out-of-distribution detection performance using synthetic outlier exposure generated by visual foundation models. In *British Machine Vision Conference (BMVC)*, 2023. 2

[40] Kimin Lee, Kibok Lee, Honglak Lee, and Jinwoo Shin. A simple unified framework for detecting out-of-distribution samples and adversarial attacks. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2018. 5, 8

[41] Hengzhuang Li and Teng Zhang. Outlier synthesis via hamiltonian monte carlo for out-of-distribution detection. In *International Conference on Learning Representations (ICLR)*, 2025. 8

[42] Jingyao Li, Pengguang Chen, Zexin He, Shaozuo Yu, Shu Liu, and Jiaya Jia. Rethinking out-of-distribution (ood) detection: Masked image modeling is all you need. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023. 8

[43] Tianqi Li, Guansong Pang, Xiao Bai, Wenjun Miao, and Jin Zheng. Learning transferable negative prompts for out-of-distribution detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024. 8

[44] Jiachen Liang, Ruibing Hou, Minyang Hu, Hong Chang, Shiguang Shan, and Xilin Chen. Revisiting logit distributions for reliable out-of-distribution detection. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2025. 2

[45] Shiyu Liang, Yixuan Li, and Rayadurgam Srikant. Enhancing the reliability of out-of-distribution image detection in neural networks. In *International Conference on Learning Representations (ICLR)*, 2018. 2, 4, 5, 8

[46] Jiahui Liu, Xin Wen, Shizhen Zhao, Yingxian Chen, and Xiaojuan Qi. Can ood object detectors learn from foundation models? In *European Conference on Computer Vision (ECCV)*. Springer, 2024. 2

[47] Litian Liu and Yao Qin. Fast decision boundary based out-of-distribution detector. In *International Conference on Machine Learning (ICML)*, 2024. 5

[48] Weitang Liu, Xiaoyun Wang, John D. Owens, and Yixuan Li. Energy-based out-of-distribution detection. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2020. 5, 8

[49] Xixi Liu, Yaroslava Lochman, and Christopher Zach. Gen: Pushing the limits of softmax-based out-of-distribution detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023. 5

[50] Yibing Liu, Chris Xing Tian, Haoliang Li, Lei Ma, and Shiqi Wang. Neuron activation coverage: Rethinking out-of-distribution detection and generalization. In *International Conference on Learning Representations (ICLR)*, 2024. 8

[51] Ziwei Liu, Ping Luo, Xiaogang Wang, and Xiaoou Tang. Deep learning face attributes in the wild. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2015. 6, 3

[52] Fan Lu, Kai Zhu, Wei Zhai, Kecheng Zheng, and Yang Cao. Uncertainty-aware optimal transport for semantically coherent out-of-distribution detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023. 8

[53] Haodong Lu, Dong Gong, Shuo Wang, Jason Xue, Lina Yao, and Kristen Moore. Learning with mixture of prototypes for out-of-distribution detection. In *International Conference on Learning Representations (ICLR)*, 2024. 8

[54] Scott M Lundberg and Su-In Lee. A unified approach to interpreting model predictions. *Advances in Neural Information Processing Systems (NeurIPS)*, 30, 2017. 3

[55] Subhransu Maji, Esa Rahtu, Juho Kannala, Matthew Blaschko, and Andrea Vedaldi. Fine-grained visual classi-

fication of aircraft. *arXiv preprint arXiv:1306.5151*, 2013. 6

[56] Yifei Ming, Hang Yin, and Yixuan Li. On the impact of spurious correlation for out-of-distribution detection. In *Proceedings of the AAAI conference on artificial intelligence*, 2022. 1, 2, 8, 3

[57] Yifei Ming, Yiyou Sun, Ousmane Dia, and Yixuan Li. How to exploit hyperspherical embeddings for out-of-distribution detection? In *International Conference on Learning Representations (ICLR)*, 2023. 5, 8

[58] Yuval Netzer, Tao Wang, Adam Coates, Alessandro Bissacco, Bo Wu, and Andrew Y Ng. Reading digits in natural images with unsupervised feature learning. 2011. 5

[59] Anh Nguyen, Jason Yosinski, and Jeff Clune. Deep neural networks are easily fooled: High confidence predictions for unrecognizable images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015. 8

[60] Jun Nie, Yadan Luo, Shanshan Ye, Yonggang Zhang, Xinmei Tian, and Zhen Fang. Out-of-distribution detection with virtual outlier smoothing. *International Journal of Computer Vision (IJCV)*, 2024. 8

[61] Bartłomiej Olber, Krystian Radlak, Adam Popowicz, Michal Szczepankiewicz, and Krystian Chachuła. Detection of out-of-distribution samples using binary neuron activation patterns. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023. 8

[62] Jaewoo Park, Jacky Chen Long Chai, Jaeho Yoon, and Andrew Beng Jin Teoh. Understanding the feature norm for out-of-distribution detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2023. 7

[63] Jaewoo Park, Yoon Gyo Jung, and Andrew Beng Jin Teoh. Nearest neighbor guidance for out-of-distribution detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2023. 5

[64] Pramuditha Perera and Vishal M Patel. Deep transfer learning for multiple class novelty detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019. 2, 8

[65] Sudarshan Regmi, Bibek Panthi, Sakar Dotel, Prashnna K Gyawali, Danail Stoyanov, and Binod Bhattarai. T2fnorm: Train-time feature normalization for ood detection in image classification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, 2024. 4, 7, 8, 18, 19, 20

[66] Sudarshan Regmi, Bibek Panthi, Yifei Ming, Prashnna K Gyawali, Danail Stoyanov, and Binod Bhattarai. Reweightood: Loss reweighting for distance-based ood detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, 2024. 8

[67] Jie Ren, Stanislav Fort, Jeremiah Liu, Abhijit Guha Roy, Shreyas Padhy, and Balaji Lakshminarayanan. A simple fix to mahalanobis distance for improving near-ood detection. *arXiv preprint arXiv:2106.09022*, 2021. 5, 8

[68] Abhijit Guha Roy, Jie Ren, Shekoofeh Azizi, Aaron Loh, Vivek Natarajan, Basil Mustafa, Nick Pawlowski, Jan Freyberg, Yuan Liu, Zach Beaver, et al. Does your dermatology classifier know what it doesn't know? detecting the long-tail of unseen conditions. *Medical Image Analysis (MedAI)*, 2022. 8

[69] Shiori Sagawa, Pang Wei Koh, Tatsunori B. Hashimoto, and Percy Liang. Distributionally robust neural networks. In *International Conference on Learning Representations (ICLR)*, 2020. 1, 2, 6

[70] Chandramouli Shama Sastry and Sageev Oore. Detecting out-of-distribution examples with gram matrices. In *International Conference on Machine Learning (ICML)*, 2020. 5

[71] Vikash Sehwag, Mung Chiang, and Prateek Mittal. Ssd: A unified framework for self-supervised outlier detection. In *International Conference on Learning Representations (ICLR)*, 2021. 8

[72] Akito Shinohara, Kohei Fukuda, and Hiroaki Aizawa. Logit mixture outlier exposure for fine-grained out-of-distribution detection. *arXiv preprint arXiv:2509.11892*, 2025. 2

[73] Daniel Smilkov, Nikhil Thorat, Been Kim, Fernanda Viégas, and Martin Wattenberg. Smoothgrad: removing noise by adding noise. *arXiv preprint arXiv:1706.03825*, 2017. 3

[74] Yue Song, Nicu Sebe, and Wei Wang. Rankfeat: Rank-1 feature removal for out-of-distribution detection. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2022. 5

[75] Yiyou Sun and Yixuan Li. Dice: Leveraging sparsification for out-of-distribution detection. In *European Conference on Computer Vision (ECCV)*, 2022. 5

[76] Yiyou Sun, Chuan Guo, and Yixuan Li. React: Out-of-distribution detection with rectified activations. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2021. 5, 8

[77] Yiyou Sun, Yifei Ming, Xiaojin Zhu, and Yixuan Li. Out-of-distribution detection with deep nearest neighbors. *International Conference on Machine Learning (ICML)*, 2022. 8

[78] Jihoon Tack, Sangwoo Mo, Jongheon Jeong, and Jinwoo Shin. Csi: Novelty detection via contrastive learning on distributionally shifted instances. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2020. 8

[79] Leitian Tao, Xuefeng Du, Jerry Zhu, and Yixuan Li. Non-parametric outlier synthesis. In *International Conference on Learning Representations (ICLR)*, 2023. 5, 8

[80] Engkarat Techapanurak and Takayuki Okatani. Practical evaluation of out-of-distribution detection methods for image classification. *arXiv preprint arXiv:2101.02447*, 2021. 2, 8

[81] Sagar Vaze, Kai Han, Andrea Vedaldi, and Andrew Zisserman. Open-set recognition: A good closed-set classifier is all you need. In *International Conference on Learning Representations (ICLR)*, 2022. 5, 6, 22

[82] Catherine Wah, Steve Branson, Peter Welinder, Pietro Perona, and Serge Belongie. The caltech-ucsd birds-200-2011 dataset. 2011. 2

[83] Haoqi Wang, Zhizhong Li, Litong Feng, and Wayne Zhang. Vim: Out-of-distribution with virtual-logit matching. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022. 5

[84] Hualiang Wang, Yi Li, Huifeng Yao, and Xiaomeng Li. Clipn for zero-shot ood detection: Teaching clip to say no. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2023. 8

[85] Qizhou Wang, Junjie Ye, Feng Liu, Quanyu Dai, Marcus Kalander, Tongliang Liu, Jianye Hao, and Bo Han. Out-of-distribution detection with implicit outlier transformation. *arXiv preprint arXiv:2303.05033*, 2023. 2

[86] Xue Wang, Zhibo Wang, Haiqin Weng, Hengchang Guo, Zhifei Zhang, Lu Jin, Tao Wei, and Kui Ren. Counterfactual-based saliency map: Towards visual contrastive explanations for neural networks. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2023. 3

[87] Yu Wang, Junxian Mu, Hongzhi Huang, Qilong Wang, Pengfei Zhu, and Qinghua Hu. Backmix: Regularizing open set recognition by removing underlying fore-background priors. *IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI)*, 2025. 2, 8

[88] Hongxin Wei, Renchunzi Xie, Hao Cheng, Lei Feng, Bo An, and Yixuan Li. Mitigating neural network overconfidence with logit normalization. In *International Conference on Machine Learning (ICML)*, 2022. 4, 5, 7, 8

[89] Ann-Christin Woerl, Jan Disselhoff, and Michael Wand. Initialization noise in image gradients and saliency maps. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023. 3

[90] Jianxiong Xiao, James Hays, Krista A Ehinger, Aude Oliva, and Antonio Torralba. SUN database: Large-scale scene recognition from abbey to zoo. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2010. 5

[91] Kai Xu, Rongyu Chen, Gianni Franchi, and Angela Yao. Scaling for training time and post-hoc out-of-distribution detection enhancement. In *International Conference on Learning Representations (ICLR)*, 2024. 5

[92] Pingmei Xu, Krista A Ehinger, Yinda Zhang, Adam Finkelstein, Sanjeev R Kulkarni, and Jianxiong Xiao. Turkergaze: Crowdsourcing saliency with webcam based eye tracking. *arXiv preprint arXiv:1504.06755*, 2015. 5

[93] Jingkang Yang, Haoqi Wang, Litong Feng, Xiaopeng Yan, Huabin Zheng, Wayne Zhang, and Ziwei Liu. Semantically coherent out-of-distribution detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2021. 8

[94] Jingkang Yang, Pengyun Wang, Dejian Zou, Zitang Zhou, Kunyuan Ding, Wenxuan Peng, Haoqi Wang, Guangyao Chen, Bo Li, Yiyou Sun, Xuefeng Du, Kaiyang Zhou, Wayne Zhang, Dan Hendrycks, Yixuan Li, and Ziwei Liu. OpenOOD: Benchmarking generalized out-of-distribution detection. In *Advances in Neural Information Processing Systems (NeurIPS), Datasets and Benchmarks Track*, 2022. 5, 8

[95] Jingkang Yang, Kaiyang Zhou, and Ziwei Liu. Full-spectrum out-of-distribution detection. *International Journal of Computer Vision (IJCV)*, 2023. 8

[96] Shaokang Yang, Shuai Liu, Cheng Yang, and Changhu Wang. Re-rank coarse classification with local region enhanced features for fine-grained image recognition. *arXiv preprint arXiv:2102.09875*, 2021. 1

[97] Yifeng Yang, Lin Zhu, Zewen Sun, Hengyu Liu, Qinying Gu, and Nanyang Ye. Oodd: Test-time out-of-distribution detection with dynamic dictionary. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2025. 2

[98] Haiyun Yao, Zongbo Han, Huazhu Fu, Xi Peng, Qinghua Hu, and Changqing Zhang. Out-of-distribution detection with diversification (provably). In *Advances in Neural Information Processing Systems (NeurIPS)*, 2024. 2, 8

[99] Qing Yu and Kiyoharu Aizawa. Unsupervised out-of-distribution detection by maximum classifier discrepancy. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2019. 8

[100] Yeonguk Yu, Sungho Shin, Seongju Lee, Changhyun Jun, and Kyoobin Lee. Block selection method for using feature norm in out-of-distribution detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023. 8

[101] Jinsong Zhang, Qiang Fu, Xu Chen, Lun Du, Zelin Li, Gang Wang, Xiaoguang Liu, Shi Han, and Dongmei Zhang. Out-of-distribution detection based on in-distribution data patterns memorization with modern hopfield energy. In *International Conference on Learning Representations (ICLR)*, 2023. 5, 8

[102] Jingyang Zhang, Nathan Inkawhich, Randolph Linderman, Yiran Chen, and Hai Li. Mixture outlier exposure: Towards out-of-distribution detection in fine-grained environments. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, 2023. 2, 5, 8

[103] Jingyang Zhang, Jingkang Yang, Pengyun Wang, Haoqi Wang, Yueqian Lin, Haoran Zhang, Yiyou Sun, Xuefeng Du, Yixuan Li, Ziwei Liu, Yiran Chen, and Hai Li. Openood v1.5: Enhanced benchmark for out-of-distribution detection. *arXiv preprint arXiv:2306.09301*, 2023. 5, 8

[104] Lily H Zhang and Rajesh Ranganath. Robustness to spurious correlations improves semantic out-of-distribution detection. In *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI)*, 2023. 2, 8

[105] Yabin Zhang, Wenjie Zhu, Chenhang He, and Lei Zhang. Lapt: Label-driven automated prompt tuning for ood detection with vision-language models. In *European Conference on Computer Vision (ECCV)*, 2025. 8

[106] Zihan Zhang and Xiang Xiang. Decoupling maxlogit for out-of-distribution detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023. 8

[107] Heliang Zheng, Jianlong Fu, Tao Mei, and Jiebo Luo. Learning multi-attention convolutional neural network for fine-grained image recognition. In *Proceedings of the*

IEEE/CVF International Conference on Computer Vision (ICCV), 2017. 1

[108] 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 (PAMI)*, 2017. 5, 2

[109] Jianing Zhu, Geng Yu, Jiangchao Yao, Tongliang Liu, Gang Niu, Masashi Sugiyama, and Bo Han. Diversified outlier exposure for out-of-distribution detection via informative extrapolation. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2023. 2, 8