SIO: Synthetic In-Distribution Data Benefits Out-of-Distribution Detection

Anonymous ICCV submission

Paper ID 2543

Abstract

Building up reliable Out-of-Distribution (OOD) detectors is challenging, often requiring the use of OOD data during training. In this work, we develop a data-driven approach which is distinct and complementary to existing works: Instead of using external OOD data, we fully exploit the internal in-distribution (ID) training set by utilizing generative models to produce additional synthetic ID images. The classifier is then trained using a novel objective that computes weighted loss on real and synthetic ID samples together. Our training framework, which is termed SIO, serves as a "plug-and-play" technique that is designed to be compatible with existing and future OOD detection algorithms, including the ones that leverage available OOD training data. Our experiments on CIFAR-10, CIFAR-100, and ImageNet variants demonstrate that SIO consistently improves the performance of nearly all state-of-the-art (SOTA) OOD detection algorithms. For instance, on the challenging CIFAR-10 v.s. CIFAR-100 detection problem, SIO improves the average OOD detection AUROC of 18 existing methods from 86.25% to 89.04% and achieves a new SOTA of 92.94% according to the OpenOOD benchmark.

1. Introduction

Being able to identify unknowns is of the utmost importance for intelligent systems to reliably operate in open world settings. In the domain of image classification, this challenge is known as *Out-of-Distribution (OOD) Detection*. The goal of OOD detection is to enable the classifier to identify samples that do not belong to one of the known, in-distribution (ID) categories during inference. However, OOD detection has long been a difficult task due to the implicit closed-world assumption adopted by standard neural network training. For instance, without certain efforts, a basic CIFAR-10 classifier will confidently predict SVHN digits as one of its classes [13].

Recent years have seen plenty of OOD detection works, which we roughly divide into three categories. 1) *Inference techniques* study post-hoc scoring rules that best separate ID and OOD data with a pre-trained model(the score indicates each sample's "OOD-ness") [2, 15, 11, 31, 32, 25, 38, 14, 14, 44, 40, 39]. 2) *Specialized training algorithms* induce more suitable structure [22, 45, 49] or feature distribution [7] inside the model via training to allow for better OOD detection at inference time. 3) *Data-driven methods* utilize additional input data for improvements. Particularly, existing works incorporate external OOD samples to either let the model learn OOD detection in a supervised way [17] or mix them with ID images as a data augmentation [19, 50]. Not surprisingly, data-driven ones are among the most effective methods as they explicitly bring in additional information outside the ID space.

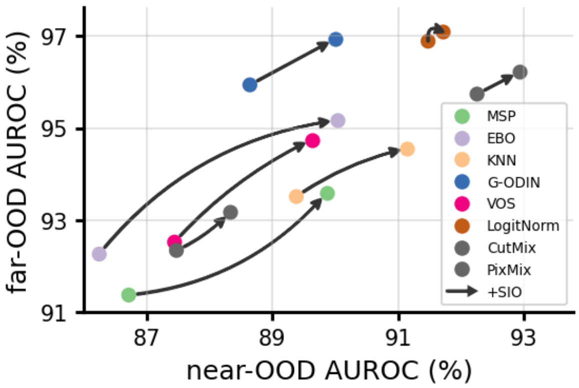


Figure 1. Results on CIFAR-10. SIO is able to yield improvements on top of multiple methods against both near- and far-OOD. See full results in Section 4.1 Table 1.

In this work, we seek to further improve OOD detection by following along the data-driven path. However, we think in an entirely different and orthogonal direction to existing approaches: While external OOD data could be helpful, *we wonder if the original, internal ID data can be better exploited to benefit OOD detection.* To answer this question, we propose using generative models (*e.g.*, GANs [28, 3], diffusion models [21]) to capture the ID data distribution and to produce additional synthetic ID images to augment the existing real ID set.

On the face of it, one may think that synthetic ID images would not benefit OOD detection and could even have a negative impact for two reasons. First, since synthetic ID images do not provide any explicit information about the OOD samples, they may not seem useful for improving OOD detection. Second, prior studies [37] have shown that naively adding synthetic images into the real training set can lower classification accuracy, which in turn may lead to inferior OOD detection performance according to the correlation between OOD detection rate and ID accuracy [43].

Contrary to initial assumptions, we reveal that under a reasonable weighting scheme, synthetic ID images can indeed be advantageous when integrated with real ID images. Concretely, we propose a novel and generic training framework which employs a weighted sum of the loss computed on both real and synthetic ID images to train the classifier. We term this framework SIO (Synthetic ID data for OOD detection). The proposed SIO framework is distinct and complementary to existing OOD detection methodologies. It can be seen as a “plug-and-play” technique that can be easily integrated into most (if not all) dedicated OOD detection approaches, including those that incorporate OOD samples in the training phase (see Section 3.2 for detailed discussion). Practically, we design a lightweight implementation of SIO that only modifies the ID batch sampling. This guarantees that SIO does not introduce any computational overhead compared to real data-only training, facilitating efficient use and fair comparison.

We conduct a thorough evaluation of SIO in combination with a variety of state-of-the-art OOD detection methods on two widely-used image classification datasets, CIFAR-10 and CIFAR-100, using the OpenOOD benchmark [46]. Our results indicate that SIO consistently improves OOD detection performance on top of multiple advanced OOD methods (see Figure 1 for some of the results on CIFAR-10). Notably, by applying SIO we achieve new state-of-the-art results in 3 out of the 4 evaluation settings within the OpenOOD benchmark. To demonstrate scalability to high-resolution images, we further experiment with two ImageNet variants. Additionally, we conduct extensive analyses to evaluate the robustness of SIO against hyperparameters, such as the choice of generative models. Finally, we observe that SIO improves OOD detection rates even if it does not improve ID classification accuracy, providing a more comprehensive view of the OOD detection rate vs. ID accuracy correlation [43].

2. Background

2.1. Problem statement

In this work we consider OOD detection in the context of multi-class image classification, where the goals are 1) training a base classifier that can accurately classify ID data and 2) building an OOD detector on top of the base classifier that accurately distinguishes OOD from ID samples. We now formulate this problem to facilitate the discussion.

Training. There are in general two types of training schemes adopted by existing works. The first one trains the base classifier f only on the labeled (real) ID training data sampled from the in-distribution set $\mathcal{D}_{\text{in}}$:

$$\min_f \mathbb{E}_{(\boldsymbol{x},y)\sim\mathcal{D}_{\text{in}}} L(\boldsymbol{x},y;f). \tag{1}$$

Here, Equation 1 is a high-level abstraction of the true optimization objectives used in practice. For instance, L could be as simple as the standard cross-entropy loss, *i.e.*, $L(\boldsymbol{x},y;f) = H(y,\sigma(f(\boldsymbol{x})))$, where $f(\boldsymbol{x})$ is the logit vector, σ is the softmax function, and y is the one-hot representation of the ground-truth label. L could also involve additional regularization as in [7, 45].

The other type of training assumes the availability of an external set of unlabeled OOD/outlier samples $\mathcal{D}_{\text{out}}$ and trains the classifier with ID and OOD samples together:

$$\min_f \mathbb{E}_{(\boldsymbol{x},y)\sim\mathcal{D}_{\text{in}},\hat{\boldsymbol{x}}\sim\mathcal{D}_{\text{out}}} L(\boldsymbol{x},y,\hat{\boldsymbol{x}};f). \tag{2}$$

For example, the loss function of Outlier Exposure (OE) [17] is $L(\boldsymbol{x},\hat{\boldsymbol{x}},y;f) = H(y,\sigma(f(\boldsymbol{x})) + \lambda \cdot H(\mathcal{U},\sigma(f(\hat{\boldsymbol{x}}))$, where $\mathcal{U}$ is the uniform distribution across the known categories, and λ is a weighting term. Later in Section 3.2, we will demonstrate the compatibility of our proposed SIO with both of these two training schemes.

Inference. After training, the OOD detector g is built upon the base classifier f with a scoring module s:

$$g(\boldsymbol{x};f,\tau) = \begin{cases} \text{OOD}, & \text{if } s(\boldsymbol{x};f) \geq \tau \\ \text{ID}, & \text{otherwise} \end{cases}. \tag{3}$$

The scoring module s will assign a score to each sample which indicates its “OOD-ness”. Again, s is a high-level representation of the scoring mechanism. Examples of s include $s(\boldsymbol{x};f) = \max_i \sigma(f(\boldsymbol{x})_i)$ (MSP [15]) and $s(\boldsymbol{x};f) = \sum_{i=1}^{K} e^{f(\boldsymbol{x})_i}$ (EBO [32]; K is the number of ID categories). Here τ is an application-dependent threshold. The detector g is used to determine whether each incoming sample is ID or OOD. For OOD samples, the base classifier will refrain from making any predictions.

2.2. Related works

1) OOD detection methodologies. There have been many works on OOD detection since it emerged as a research problem. We refer readers to [47] for a comprehensive survey, while we focus on several top-performing methods that we consider in this work. As aforementioned, we roughly categorize existing works into three groups.

The first line of works focus on the design of the post-hoc scoring rule s (Equation 3), while assuming that the base classifier is pre-trained (usually with the standard cross-entropy loss). In general, the outputs from the model's decision space (*e.g.*, the final linear layer or the penultimate layer) could derive informative scores. For example, MSP [15] and MLS [14] directly use the (negative) maximum softmax probability and maximum logit value as the score, respectively. Other works post-process the network's outputs to enlarge the difference between ID and OOD. Examples include softening softmax probabilities with temperature scaling [11], applying energy function to the logits [32], rectifying activations with thresholding [38], and applying KNN to the penultimate layer's features [40], etc.

While some researchers focus on inference techniques, others investigate specialized training algorithms that involve developing more advanced L in Equation 1. Notably, G-ODIN [22] trains a dividend/divisor structure to decompose the softmax confidence. VOS [7] encourages the learned representations of the classifier to shape like a mixture of Gaussian distributions. LogitNorm [45] trains the model with normalized logit vectors as the unit sphere provides more discriminative information for distinguishing ID and OOD samples. Meanwhile, using certain tricks such as longer training, stronger data augmentation, and more complex learning rate schedule has also been shown to benefit OOD detection [43].

The final group of works proposes data-driven approaches to improve OOD detection. In particular, OE [17] collects real OOD samples and explicitly lets the model learn OOD detection in a supervised manner. PixMix [19] applies pixel-level mixing operations between ID images and low-level OOD images (which exhibit certain visual patterns but do not have clear semantics) as an augmentation. Additionally, pre-training with a huge labeled dataset has been shown to be beneficial for OOD detection in downstream tasks (*e.g.*, using ImageNet pre-training for CIFAR-10) [16]. *It is noteworthy that all these data-driven methods require external data beyond the original ID training set.*

While SIO, like other data-driven approaches, introduces additional input images, it stands out from existing methods by utilizing synthetic ID images instead of external OOD images. This approach is particularly useful in data-scarce scenarios, where external data may not be readily available. Furthermore, SIO is inherently complementary to all previously discussed methods, including data-driven ones. We provide detailed discussions and empirical evidence to support this claim in the subsequent sections.

2) Synthetic images for OOD detection. Although previous studies in OOD detection research have used synthetic images produced by generative models, they have all focused on synthesizing OOD images [8, 35, 30]. However, synthesizing OOD images is a challenging task as there could be a lack of real OOD data for supervision, and the distribution of the open space is too broad to capture. As a result, previous attempts have not yielded superior performance, with models trained on synthetic OOD images underperforming those trained on real OOD data [17] or even simple ID-only training baselines [43].

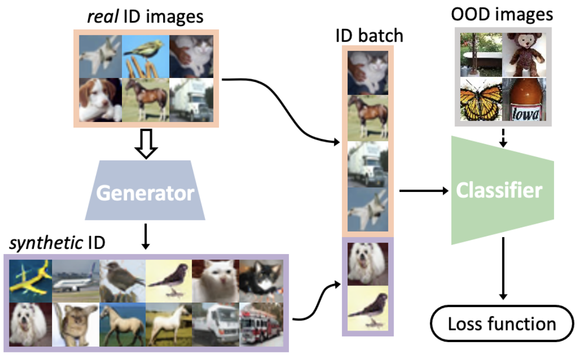


Figure 2. Overview of the proposed SIO framework. First, a large number of synthetic ID samples are generated (offline) using a generative model that is trained with real ID images. Then, we combine the real and synthetic ID samples to train the classifier in a weighted fashion described in Equation 4. The specific loss function can be selected from any existing OOD training methods. External OOD samples can also be incorporated into the training depending on their availability (Equation 5).

In sharp contrast, what we propose is using generative models to synthesize ID images, which is much easier a task since the target distribution is well-defined (characterized by the ID training set). In addition, we demonstrate that incorporating synthetic ID samples can significantly improve the performance of multiple OOD detection methods.

3) Synthetic samples for other performance measures. Expanding the real training set with synthetic samples has been demonstrated to help with adversarial training [10], where overfitting is the primary challenge that hinders better adversarial robustness. In standard, non-adversarial settings, however, including synthetic samples has not yet led to improvements in, *e.g.*, classification accuracy [37]. To the best of our knowledge, we are the first to investigate and demonstrate that synthetic ID images can enhance OOD detection. Our study provides a new avenue for exploring the use of synthetic data in other areas of machine learning.

3. Methodology

We now describe the proposed SIO as a two-step procedure, including several technical details. See Figure 2 for an overview.

3.1. Generating synthetic ID training set $\mathcal{D}_{\text{in}}^{\text{syn}}$

Given the real ID training set $\mathcal{D}_{\text{in}}^{\text{real}}$, we train a generative model to capture the ID data distribution, which can then be used to generate a large number of synthetic ID images by sampling from the synthetic distribution $\mathcal{D}_{\text{in}}^{\text{syn}}$. The training step can be skipped if pre-trained generative models are available. For instance, we utilize off-the-shelf generative models that are pre-trained on standard ID datasets (*e.g.*, CIFAR-10) in most of our experiments. It is worth noting that a generator trained on a different (potentially larger) dataset can also be used, as long as it knows the concept of ID categories (*e.g.*, an ImageNet generator can be employed for a Tiny ImageNet classifier).

The way the labels of synthetic samples are obtained can vary depending on the case. If the real ID dataset $\mathcal{D}_{\text{in}}^{\text{real}}$ has labels (which is typically the case), the generative model can be trained in a class-conditional way, and we can directly sample labeled synthetic data $(\tilde{\boldsymbol{x}}, \tilde{y}) \sim \mathcal{D}_{\text{in}}^{\text{syn}}$. If $\mathcal{D}_{\text{in}}^{\text{real}}$ is unlabeled (*e.g.*, in a self-supervised setting) or the generative model is unconditional, we can only sample unlabeled synthetic data $\tilde{\boldsymbol{x}} \sim \mathcal{D}_{\text{in}}^{\text{syn}}$. In this case, we use a classifier f pre-trained on the real data to produce pseudo-labels, i.e., $\tilde{y} = \arg\max_i f(\tilde{\boldsymbol{x}})_i$. By default, we assume that the generative model is conditional unless otherwise stated.

In this work, we consider the generation step as offline, meaning that we pre-generate a fixed number of synthetic images. Therefore, it does not add any complexity to the online training step. However, if there are sufficient computing resources, synthetic samples can certainly be generated on the fly. In fact, our experiments indicate that more synthetic samples generally lead to better performance.

3.2. Training with real and synthetic ID samples

Subtle differences or shifts can exist between the synthetic and real distributions [37]. To avoid bias towards the synthetic distribution, we train the model using real and synthetic data together. Concretely, we design a weighted objective, which is essential for obtaining superior performance as shown in later experiments:

$$\min_f \Big[\alpha \cdot \mathbb{E}_{(\boldsymbol{x},y)\sim\mathcal{D}_{\text{in}}^{\text{real}}} L(\boldsymbol{x}, y; f) + (1-\alpha) \cdot \mathbb{E}_{(\tilde{\boldsymbol{x}},\tilde{y})\sim\mathcal{D}_{\text{in}}^{\text{syn}}} L(\tilde{\boldsymbol{x}}, \tilde{y}; f)\Big], \quad (4)$$

where $\alpha \in [0, 1]$ is the weighting term. Note that, similarly to Equation 1, L could be any specific loss function of existing OOD training algorithms. SIO can also be seamlessly integrated with methods that incorporate OOD samples during the training (Equation 2). In such cases, the objective is:

$$\min_f \Big[\alpha \cdot \mathbb{E}_{(\boldsymbol{x},y)\sim\mathcal{D}_{\text{in}}^{\text{real}}, \hat{\boldsymbol{x}}\sim\mathcal{D}_{\text{out}}} L(\boldsymbol{x}, y, \hat{\boldsymbol{x}}; f) + (1-\alpha) \cdot \mathbb{E}_{(\tilde{\boldsymbol{x}},\tilde{y})\sim\mathcal{D}_{\text{in}}^{\text{syn}}, \hat{\boldsymbol{x}}\sim\mathcal{D}_{\text{out}}} L(\tilde{\boldsymbol{x}}, \tilde{y}, \hat{\boldsymbol{x}}; f)\Big]. \quad (5)$$

In practice, SIO is implemented in an equivalent but much more efficient way than its basic form in Equation 4 and 5. Instead of performing two separate forward passes and computing the weighted loss, we do the weighting inside each mini-batch by replacing a certain amount of real ID samples with synthetic ones such that the ratio between real and synthetic ID samples is $\alpha : 1-\alpha$ within each batch (see Figure 2). Such implementation avoids additional computation overhead and allows fair comparison to real data-only training since each model is trained under the exact same budget.

4. Experiments

We demonstrate the effectiveness of additional synthetic ID data for OOD detection through extensive experiments and analyses, starting with results in settings where CIFAR-10/100 [29] is used as the ID dataset. We then scale up to high-resolution settings with ImageNet splits. Finally, we analyze SIO's robustness to hyperparameters.

4.1. CIFAR

Benchmarking OOD detection methods used to be challenging because there was no unified platform with standardized implementations and benchmarks, making it difficult to make direct comparisons. However, the recent work called OpenOOD [46] provides such a platform that enables fair and accurate benchmarking. In this work, we use OpenOOD's setup as a basis for our experiments while making some modifications where necessary.

Baselines. Since SIO is orthogonal to existing OOD detection methodologies, we demonstrate the effectiveness of SIO by combining it with multiple OOD detection approaches and comparing the results with the real data-only counterpart in each case. Specifically, we consider 11 inference techniques [2, 15, 11, 31, 32, 38, 14, 14, 44, 40, 39], 5 specialized training algorithms [22, 45, 49, 7, 43], and 2 data-driven methods [19, 17], resulting in a total of 18 OOD methods. All of them are the top-performing ones according to the OpenOOD benchmark.

Training setup. We use ResNet-18 [12] as the classifier architecture, following OpenOOD. Regardless of the training algorithm, we train the model using Nesterov SGD with a momentum of 0.9. The initial learning rate is set to 0.1 and is decayed according to the cosine annealing schedule [33]. A weight decay of 0.0005 is applied during training, and the batch size is set to 128. The only deviation from the OpenOOD default configuration is that we train each model for 200 epochs instead of 100 epochs, as longer training has been shown to improve OOD performance [43]. For method-specific hyperparameters, we use the default or recommended values provided by OpenOOD.

For our proposed SIO, we generate 1000K synthetic ID samples using StyleGAN2 [28] for both the CIFAR-10 and CIFAR-100 datasets. The weighting/ratio factor α (Equation 4 and 5) is set to 0.8, *i.e.*, there will be 20% synthetic ID samples in each ID training batch. We remark that the SIO model and the real data-only model undergo the exact same number of gradient steps and share the same batch

Table 1. Results in terms of detection AUROC (%) on CIFAR-10/100 datasets. All numbers are percentages and are averaged over 3 runs. In most cases introducing synthetic ID data leads to improvements. Note that these numbers cannot be directly compared with those reported in the OpenOOD paper [46] due to differences in training and evaluation setup. See text for detailed description.

Method	CIFAR-10						CIFAR-100					
	near-OOD		far-OOD		ID Accuracy		near-OOD		far-OOD		ID Accuracy	
	baseline	+SIO	baseline	+SIO	baseline	+SIO	baseline	+SIO	baseline	+SIO	baseline	+SIO
- Inference techniques												
OpenMax [2]	85.14	87.99	90.08	92.50	94.93	95.56	73.78	75.12	70.33	73.56	78.06	78.67
MSP [15]	86.71	89.86	91.37	93.59			79.56	79.04	80.10	82.12		
TempScale [11]	86.30	89.72	92.39	94.57			79.54	79.38	83.93	84.71		
ODIN [31]	77.86	85.71	86.68	92.62			79.12	78.89	81.02	84.23		
EBO [32]	86.24	90.03	92.26	95.17			80.17	80.20	81.51	85.50		
ReAct [38]	84.63	89.29	90.86	94.40			72.52	69.45	78.33	85.84		
MLS [14]	86.17	89.99	92.11	95.00			80.22	80.34	81.33	85.25		
KLM [14]	79.85	81.50	85.41	87.57			73.38	75.12	76.17	80.32		
VIM [44]	84.16	88.23	90.43	92.51			63.70	72.00	74.06	82.89		
KNN [40]	89.38	91.15	93.51	94.56			77.75	77.65	82.34	86.13		
DICE [39]	80.20	85.54	87.78	92.56			80.06	79.86	81.63	84.71		
- Training algorithms												
G-ODIN [22]	88.64	90.00	95.94	96.94	94.89	95.48	72.50	73.83	87.54	89.37	75.10	77.00
VOS [7]	87.44	89.63	92.52	94.75	95.20	95.62	80.05	79.32	81.12	84.39	78.37	78.83
LogitNorm [45]	91.48	91.72	96.89	97.09	94.52	95.27	74.96	75.48	82.60	85.81	76.16	78.05
CutMix [49]	87.47	88.33	92.34	93.18	96.48	96.48	78.19	78.86	76.61	79.18	80.41	81.23
CrossEntropy+ [43]	89.99	91.00	93.42	93.81	95.84	95.87	78.92	79.47	78.77	80.26	77.62	79.29
- Data-driven methods												
PixMix [19]	92.26	92.94	95.74	96.22	95.47	96.01	76.93	78.13	84.76	85.99	77.82	79.61
OE [17]	88.62	90.16	97.16	96.92	95.17	95.59	78.00	77.40	83.01	85.43	77.81	77.99
Average	86.25	89.04	92.05	94.11	95.10	95.64	76.63	77.20	80.29	83.65	77.89	78.74
Best	92.26	92.94	97.16	97.09	96.48	96.48	80.22	80.34	87.54	89.37	80.41	81.23

size, ensuring a fair comparison. The only difference is that the SIO model sees more diverse training samples by leveraging synthetic ID data.

Evaluation setup. For CIFAR-10/100, we consider CIFAR-100/10 as near-OOD and MNIST [6], SVHN [36], Texture [4], and Places365 [51] as far-OOD. Far-OOD samples are semantically far away from the ID samples and meanwhile often exhibit significant low-level, non-semantic shifts as well (due to data collection differences) [1, 41], making them easier to detect. Near-OOD samples are more similar to ID samples in both semantic and non-semantic aspects and are more challenging to identify.

We follow the OpenOOD setting with one exception: We remove Tiny ImageNet samples from the near-OOD split since they are used as the training OOD samples for OE [17]. If not removed, the training and test OOD distributions would completely overlap, resulting in a trivial case.

We use the area under the receiver operating characteristic curve (AUROC) as the metric for evaluation. AUROC is a threshold-independent metric for binary classification; the higher the better, and the random-guessing baseline is 50%. We report the detection AUROC against near- and far-OOD (averaged over the OOD sets in each split) for each method. We also report the classification accuracy on ID test data, as a good algorithm should not trade-off ID accuracy for OOD detection performance. All reported results are averaged over 3 independent training runs.

Results. See Table 1. We start with discussing CIFAR-10 results. The first takeaway is that SIO works well with nearly all OOD methods and helps with both near- and far-OOD detection, leading to noticeable improvements in 35 out of 36 cases (18 methods × {near-OOD, far-OOD}). On average, SIO improves the near-OOD and far-OOD AUROC from 86.25% to 89.04% and from 92.05% to 94.11%, respectively.

Interestingly, we find that the performance gains brought by SIO can sometimes surpass those achieved through dedicated algorithmic design. For example, SIO improves the near-OOD / far-OOD AUROC of the MSP detector from 86.71% / 91.37% to 89.86% / 93.59%, outperforming the more complex KNN detector on the baseline model, which achieves 89.38% / 93.51% at the cost of significantly increased inference latency [34]. Furthermore, we confirm that SIO is compatible with data-driven methods that incorporate external OOD data (OE and PixMix), suggesting that the information contained in synthetic ID samples is complementary to that in external OOD samples.

Our second takeaway is that SIO improves the state-of-the-art result on the challenging CIFAR-10 near-OOD detection from 92.26% to 92.94%. On CIFAR-10 far-OOD

Table 2. Results in terms of detection AUROC (%) on two ImageNet variants. All numbers are percentages and are averaged over 3 runs. \ means that the NaN error occurs when evaluating that method.

Method	ImageNet-10						ImageNet-dogs					
	near-OOD		far-OOD		ID Accuracy		near-OOD		far-OOD		ID Accuracy	
	baseline	+SIO	baseline	+SIO	baseline	+SIO	baseline	+SIO	baseline	+SIO	baseline	+SIO
OpenMax [2]	84.39	86.19	89.14	90.53	90.00	91.78	95.68	96.01	93.70	92.96	75.42	75.34
MSP [15]	84.27	87.25	88.52	92.03			95.81	96.17	94.58	95.80		
TempScale [11]	84.92	87.75	89.24	92.70			97.22	97.37	96.51	97.17		
ODIN [31]	87.30	89.14	93.31	94.80			97.19	97.39	98.41	98.63		
EBO [32]	86.09	88.18	91.74	93.93			97.96	98.12	98.51	98.80		
GradNorm [25]	87.23	87.67	95.86	94.71			95.54	96.05	98.32	99.27		
ReAct [38]	86.21	87.61	92.37	94.50			97.98	98.13	99.21	99.47		
MLS [14]	86.26	88.36	91.68	93.92			97.91	98.05	98.30	98.62		
KLM [14]	77.07	80.52	83.65	89.19			\	\	\	\		
VIM [44]	\	\	\	\			\	\	\	\		
KNN [40]	87.81	89.73	97.79	97.65			98.11	98.26	99.84	99.86		
DICE [39]	87.29	89.81	96.06	95.92			97.26	97.64	98.64	99.18		
Average	85.35	87.47	91.76	93.63	90.00	91.78	97.07	97.32	97.60	97.98	75.42	75.34
Best	87.81	89.81	97.79	97.65	90.00	91.78	98.11	98.26	99.84	99.86	75.42	75.34

detection, SIO does not outperform vanilla OE, which incorporates a large set of external OOD samples for training. We suspect that this is because far-OOD detection on CIFAR-10 is dominated by low-level statistics [1, 41], and using synthetic ID samples may slightly push the model away from the real low-level statistics of the ID data, causing a shrinked difference between ID and OOD samples. Nonetheless, SIO improves far-OOD detection in all other cases and effectively closes the gap between non-OE methods and OE. Notably, LogitNorm + SIO yields a 97.09% AUROC without any OOD training data, which is on par with the 97.16% AUROC achieved by OE.

Lastly, we observe that SIO can also benefit ID classification accuracy. While this is not the focus of this work, our finding suggests that synthetic samples can be helpful for accuracy if used properly, challenging previous beliefs [37]. On the other hand, however, in our later experiments where we vary the hyperparameters, we find that SIO consistently boosts OOD detection performance even if it does not improve ID accuracy. More discussion on this can be found in Section 4.3.

On CIFAR-100, we observe similar results to CIFAR-10, with the general observation that SIO can benefit OOD detection performance. With SIO, the average near-OOD / far-OOD AUROC is lifted from 76.63% / 80.29% to 77.20% / 83.65%, and the best numbers are improved from 80.22% / 87.54% to 80.34% / 89.37%.

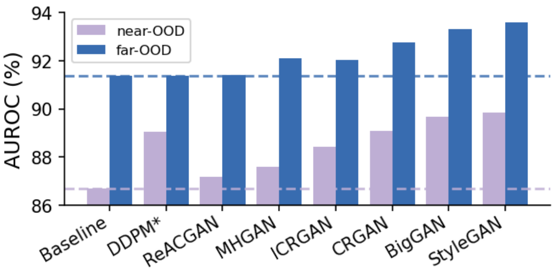
4.2. Scaling to high-resolution images

We now investigate whether SIO's effect can extend to high-resolution images at the scale of ImageNet. However, we note that generative modeling on certain ImageNet categories remains a challenging problem due to inherent difficulties in the data [3, 37]. For example, the category `tench` includes many images depicting the fish being held by human beings, which causes the generative model to learn irrelevant information unrelated to the target object itself. Consequently, generated images may appear unnatural and deviate significantly from the true distribution (see Appendix A Figure 7 for visual examples). In such cases, including unrealistic synthetic images during training is unlikely to be beneficial. This issue, however, is related to generative modeling rather than our proposed SIO and is expected to be mitigated as generative modeling techniques continue to advance. For our experimental purpose, here we utilize two subsets of ImageNet categories where plausible images can be generated. This allows us to examine the effectiveness of SIO in high-resolution settings with limited generative modeling challenges.

Datasets. The first subset is ImageNet-10, which is similar to CIFAR-10 with categories such as `aircraft`, `automobile`, and `bird`. The other one is ImageNet-dogs [24], which consists of 100 dog categories. See Appendix A for the complete list of class WordNet IDs.

Training setup. We train ResNet-18 models for 60 epochs using SGD with a momentum of 0.9. The initial learning rate is 0.1 and decays according to the cosine annealing schedule. We use a batch size of 256 and the standard `RandomResizedCrop` with the final size being 224x224 as data augmentation. For our SIO, we generate 10K synthetic images per category for both ImageNet-10 and ImageNet-dogs datasets using BigGAN [3]. The weighting/ratio factor α in Equation 4 is set to 0.9.

Evaluation setup. Due to the high (sometimes unaffordable) computational cost of many specialized training methods [43, 46], we only evaluate inference techniques



data source	FID↓	AUROC↑
DDPM*	3.17	90.22
ReACGAN	4.40	89.29
MHGAN	4.18	89.84
ICRGAN	3.52	90.24
CRGAN	3.48	90.94
BigGAN	2.98	91.49
StyleGAN	2.67	91.73

Figure 3. **Left:** OOD detection results of using different generative models. SIO is fairly robust to the choice of generative model. **Right:** We find that the FID metric of synthetic images correlates well with the corresponding OOD detection performance (the average of near- and far-OOD AUROC). *Unlike GAN models, the diffusion model (DDPM) here is unconditional.

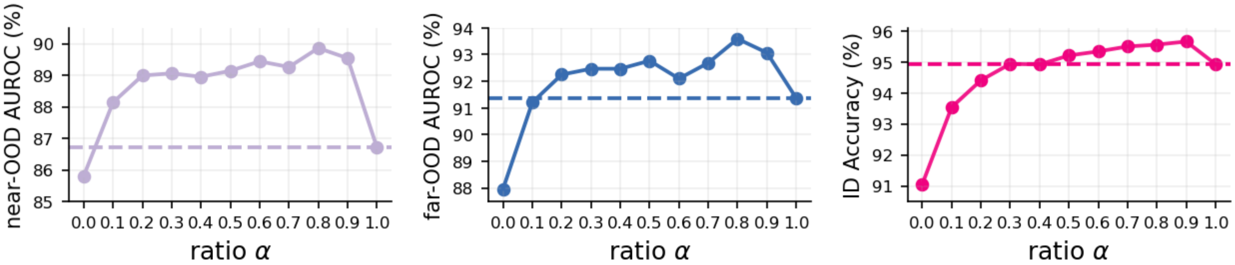


Figure 4. Results of varying the ratio α in Equation 4. Compared with the real data-only baseline (α=1.0), SIO consistently leads to improvements in a wide range of α (from 0.2 to 0.9).

with standard cross-entropy training, following OpenOOD. However, it should be noted that standard training is in fact a strong baseline on the ImageNet scale [43].

For ImageNet-10, we use Species [14], iNaturalist [42], ImageNet-O [18], and OpenImage-O [44] as near-OOD, as per OpenOOD. For ImageNet-dogs, we take non-dog ImageNet samples as near-OOD, which is a more challenging problem [24]. For both datasets, Textures [4], MNIST [6], and SVHN [36] are used as far-OOD [46]. In line with our previous experiments, we report the average near-OOD / far-OOD AUROC and ID accuracy from 3 runs.

Results. Our results on the ImageNet splits demonstrate that the benefits of SIO observed in CIFAR experiments generalize to high-resolution images. On ImageNet-10, while SIO does not achieve a higher best AUROC against far-OOD, it does improve the average near-OOD / far-OOD AUROC from 85.35% / 91.76% to 87.47% / 93.63% and achieves the best near-OOD AUROC of 89.81% compared to the best score of 87.81% from real data-only baselines. On ImageNet-dogs, interestingly, SIO slightly degrades the ID accuracy, but still yields better average and best results than the real data-only baselines.

4.3. Analyses

To analyze SIO's robustness to its hyperparameters, we conduct several experiments on CIFAR-10 using MSP as the detector.

Choice of the generative model. In our CIFAR experiments, we used StyleGAN to generate synthetic data. We now investigate whether SIO remains effective when using other generative models by repeating the SIO training with several other class-conditional GANs (using pre-trained models from [27]). We also consider an unconditional diffusion model [21], where we use pseudo-labels for synthetic images, as discussed in Section 3.1. The results are shown in the left panel of Figure 3. We find that in all cases, SIO yields noticeable performance gains over the real data-only baseline, demonstrating that SIO is effective regardless of the specific choice of the generative model.

However, we also observe that certain generative models outperform others when used as the source of synthetic data. This leads us to consider metrics that can explain relative performance and guide the selection of the generative model beforehand. Intuitively, the *quality* and *diversity* are two important considerations, meaning that we want synthetic samples to be realistic and diverse such that they can provide useful information in addition to the real data. Since advanced generative models can already provide sufficient diversity on small-scale datasets like CIFAR [28], we hypothesize that in our experiments the sample quality is the dominant factor. To measure sample quality, we use the well-established Fréchet Inception Distance (FID) [20], which measures the distance between the synthetic distribution and the real distribution in the Inception feature space. The results are shown in the right panel of Figure 3. We find that the FID metric correlates well with OOD detection performance (the average of near- and far-OOD AUROC). For example, StyleGAN images exhibit the lowest FID (best sample quality) and lead to the highest OOD detection results. The only exception to this correlation is the unconditional diffusion model, where we suspect that the quality of the pseudo-labels may also have an effect.

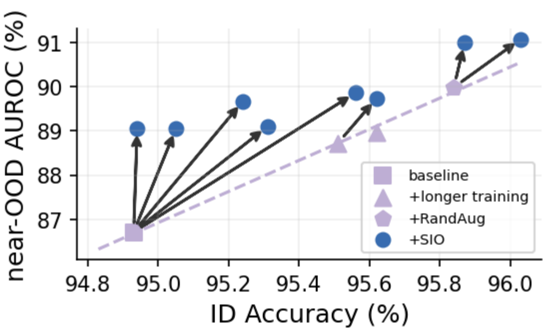


Figure 5. Near-OOD detection AUROC v.s. ID classification accuracy plot. SIO benefits OOD detection even when there are limited accuracy improvements. It also leads to a greater net gain in OOD detection per unit increase in ID accuracy, compared to the tricks (longer training and RandAug) used in [43].

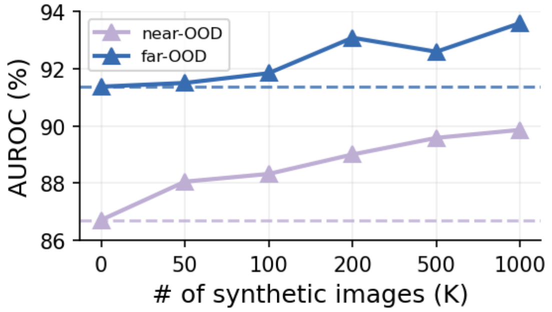


Figure 6. Results of varying the number of synthetic ID images. More samples (higher diversity) in general lead to better performance.

The ratio α. Another hyperparameter of SIO is the weighting/ratio α in Equation 4 and 5. To investigate the effect of α, we vary it from 0.0 (using only synthetic samples) to 1.0 (using only real samples) and evaluate the resulting OOD detection performance. Results are presented in Figure 4.

We identify that SIO consistently provides notable improvements in OOD detection performance across a wide range of α. The best results are obtained under a large α (*e.g*., 0.8 or 0.9), with the majority of the samples still being real data. We think this is because oversampling synthetic samples will exacerbate the distribution shift between the synthetic and real distribution, leading to worse performance. Such effect can also be observed in the ID classification accuracy, where biasing towards synthetic data (α<0.5) reduces the accuracy compared with the real data baseline. Meanwhile, notice that naively injecting synthetic images into the training set, which corresponds to $\alpha \approx 0.05$ in Figure 4 (there are 50K real samples and 1000K synthetic samples), would degrade the performance. Our results highlight that the weighting scheme adopted by SIO is the key ingredient of making synthetic images useful.

Classifier architecture. To demonstrate that SIO can work with other classifier architectures, we repeat the CIFAR-10 experiments with DenseNet-100 [23]. The experimental setup and hyperparameters remain the same as before, and we focus on inference techniques for simplicity. See Appendix B for full results. Overall, SIO improves the average scores from 84.58% / 84.59% to 86.04% / 89.50% and the best scores from 90.12% / 93.81% to 90.64% / 95.43% against near-OOD / far-OOD, respectively.

Explaining SIO's effects. It is possible to assume that SIO's ability to improve OOD detection performance is solely due to its enhancement of ID classification accuracy. While this assumption aligns with the previous finding in [43] which suggests that there is a correlation between OOD detection performance and ID accuracy, our analysis indicates that this view does *not* fully explain SIO's effects.

To demonstrate this, we reproduce the OOD-ID correlation observed in [43] by applying the techniques used in that study, including longer training and RandAug [5]. The results of this experiment, together with the results of the SIO training, are presented in Figure 5. The diversity of the blue dots comes from varying SIO's hyperparameters. We observe that SIO improves OOD detection performance even when it does not enhance ID classification accuracy. Additionally, we find that SIO yields a higher net gain in OOD detection performance per unit increase in ID accuracy than purely chasing a better classifier, as done in [43].

We hypothesize that SIO's effects stem from the increased diversity provided by additional synthetic ID samples. Our reasoning is based on the observation that neural networks often rely on "spurious" or "shortcut" features in images that are only superficially correlated with the labels [48, 26, 9]. The challenge of OOD detection, then, is that OOD samples can easily activate these spurious features. By training the model on more diverse ID samples, each potentially coming with different spurious features, the model may rely less on such features and become less likely to activate when presented with OOD samples.

To test our hypothesis, we vary the number of synthetic images as a proxy for the diversity of the training set. The results presented in Figure 6 provide supporting evidence for our hypothesis. The plot shows a generally increasing trend in OOD detection performance as the number of synthetic images increases.

5. Conclusion

In this work, we propose SIO, a training framework that utilizes synthetic ID samples and a weighted objective to benefit OOD detection. SIO can be easily integrated with existing approaches and consistently improves OOD detection performance on top of multiple OOD detectors. Importantly, our findings suggest that training with additional and diverse synthetic ID data benefits OOD detection, which may open up new directions for future research. As generative models continue to advance, we expect SIO to remain an effective and versatile component for OOD detection.

References

[1] Faruk Ahmed and Aaron Courville. Detecting semantic anomalies. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 3154–3162, 2020. 5, 6

[2] Abhijit Bendale and Terrance E Boult. Towards open set deep networks. In *CVPR*, 2016. 1, 4, 5, 6, 12

[3] Andrew Brock, Jeff Donahue, and Karen Simonyan. Large scale GAN training for high fidelity natural image synthesis. In *International Conference on Learning Representations*, 2019. 1, 6

[4] M. Cimpoi, S. Maji, I. Kokkinos, S. Mohamed, , and A. Vedaldi. Describing textures in the wild. In *Proceedings of the IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2014. 5, 7

[5] Ekin D Cubuk, Barret Zoph, Jonathon Shlens, and Quoc V Le. Randaugment: Practical automated data augmentation with a reduced search space. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition workshops*, pages 702–703, 2020. 8

[6] Li Deng. The mnist database of handwritten digit images for machine learning research. *IEEE Signal Processing Magazine*, 2012. 5, 7

[7] Xuefeng Du, Zhaoning Wang, Mu Cai, and Yixuan Li. Vos: Learning what you don't know by virtual outlier synthesis. In *ICLR*, 2022. 1, 2, 3, 4, 5

[8] ZongYuan Ge, Sergey Demyanov, Zetao Chen, and Rahil Garnavi. Generative openmax for multi-class open set classification. *arXiv preprint arXiv:1707.07418*, 2017. 3

[9] Justin Gilmer and Dan Hendrycks. A discussion of 'adversarial examples are not bugs, they are features': Adversarial example researchers need to expand what is meant by 'robustness'. *Distill*, 2019. https://distill.pub/2019/advex-bugs-discussion/response-1. 8

[10] Sven Gowal, Sylvestre-Alvise Rebuffi, Olivia Wiles, Florian Stimberg, Dan Andrei Calian, and Timothy A Mann. Improving robustness using generated data. *Advances in Neural Information Processing Systems*, 34:4218–4233, 2021. 3

[11] Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q Weinberger. On calibration of modern neural networks. In *ICML*, 2017. 1, 3, 4, 5, 6, 12

[12] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016. 4

[13] Matthias Hein, Maksym Andriushchenko, and Julian Bitterwolf. Why relu networks yield high-confidence predictions far away from the training data and how to mitigate the problem. In *CVPR*, 2019. 1

[14] Dan Hendrycks, Steven Basart, Mantas Mazeika, Mohammadreza Mostajabi, Jacob Steinhardt, and Dawn Song. Scaling out-of-distribution detection for real-world settings. In *ICML*, 2022. 1, 3, 4, 5, 6, 7, 12

[15] Dan Hendrycks and Kevin Gimpel. A baseline for detecting misclassified and out-of-distribution examples in neural networks. In *ICLR*, 2017. 1, 2, 3, 4, 5, 6, 12

[16] Dan Hendrycks, Kimin Lee, and Mantas Mazeika. Using pre-training can improve model robustness and uncertainty. In *International Conference on Machine Learning*, pages 2712–2721. PMLR, 2019. 3

[17] Dan Hendrycks, Mantas Mazeika, and Thomas Dietterich. Deep anomaly detection with outlier exposure. In *ICLR*, 2019. 1, 2, 3, 4, 5

[18] Dan Hendrycks, Kevin Zhao, Steven Basart, Jacob Steinhardt, and Dawn Song. Natural adversarial examples. *CVPR*, 2021. 7

[19] Dan Hendrycks, Andy Zou, Mantas Mazeika, Leonard Tang, Dawn Song, and Jacob Steinhardt. Pixmix: Dreamlike pictures comprehensively improve safety measures. *arXiv preprint arXiv:2112.05135*, 2021. 1, 3, 4, 5

[20] Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium. *Advances in neural information processing systems*, 30, 2017. 7

[21] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in Neural Information Processing Systems*, 33:6840–6851, 2020. 1, 7

[22] 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 *CVPR*, 2020. 1, 3, 4, 5

[23] Gao Huang, Zhuang Liu, Laurens Van Der Maaten, and Kilian Q Weinberger. Densely connected convolutional networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4700–4708, 2017. 8

[24] Haiwen Huang, Zhihan Li, Lulu Wang, Sishuo Chen, Bin Dong, and Xinyu Zhou. Feature space singularity for out-of-distribution detection. In *Proceedings of the Workshop on Artificial Intelligence Safety 2021 (SafeAI 2021)*, 2021. 6, 7

[25] Rui Huang, Andrew Geng, and Yixuan Li. On the importance of gradients for detecting distributional shifts in the wild. In *NeurIPS*, 2021. 1, 6

[26] Andrew Ilyas, Shibani Santurkar, Dimitris Tsipras, Logan Engstrom, Brandon Tran, and Aleksander Madry. Adversarial examples are not bugs, they are features. *Advances in neural information processing systems*, 32, 2019. 8

[27] Minguk Kang, Joonghyuk Shin, and Jaesik Park. Studiogan: A taxonomy and benchmark of gans for image synthesis. *arXiv preprint arXiv:2206.09479*, 2022. 7

[28] Tero Karras, Miika Aittala, Janne Hellsten, Samuli Laine, Jaakko Lehtinen, and Timo Aila. Training generative adversarial networks with limited data. *Advances in neural information processing systems*, 33:12104–12114, 2020. 1, 4, 7

[29] Alex Krizhevsky, Geoffrey Hinton, et al. Learning multiple layers of features from tiny images. 2009. 4

[30] Kimin Lee, Honglak Lee, Kibok Lee, and Jinwoo Shin. Training confidence-calibrated classifiers for detecting out-of-distribution samples. 2018. 3

[31] Shiyu Liang, Yixuan Li, and Rayadurgam Srikant. Enhancing the reliability of out-of-distribution image detection in neural networks. In *ICLR*, 2018. 1, 4, 5, 6, 12

[32] Weitang Liu, Xiaoyun Wang, John D Owens, and Yixuan Li. Energy-based out-of-distribution detection. In *NeurIPS*, 2020. 1, 2, 3, 4, 5, 6, 12

[33] Ilya Loshchilov and Frank Hutter. SGDR: Stochastic gradient descent with warm restarts. In *International Conference on Learning Representations*, 2017. 4

[34] Siyu Luan, Zonghua Gu, Amin Saremi, Leonid Freidovich, Lili Jiang, and Shaohua Wan. Timing performance benchmarking of out-of-distribution detection algorithms. *Journal of Signal Processing Systems*, 02 2023. 5

[35] Lawrence Neal, Matthew Olson, Xiaoli Fern, Weng-Keen Wong, and Fuxin Li. Open set learning with counterfactual images. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 613–628, 2018. 3

[36] 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, 7

[37] Suman Ravuri and Oriol Vinyals. Classification accuracy score for conditional generative models. *Advances in neural information processing systems*, 32, 2019. 2, 3, 4, 6

[38] Yiyou Sun, Chuan Guo, and Yixuan Li. React: Out-of-distribution detection with rectified activations. In *NeurIPS*, 2021. 1, 3, 4, 5, 6, 12

[39] Yiyou Sun and Sharon Li. Dice: Leveraging sparsification for out-of-distribution detection. In *ECCV*, 2022. 1, 4, 5, 6, 12

[40] Yiyou Sun, Yifei Ming, Xiaojin Zhu, and Yixuan Li. Out-of-distribution detection with deep nearest neighbors. *ICML*, 2022. 1, 3, 4, 5, 6, 12

[41] Jihoon Tack, Sangwoo Mo, Jongheon Jeong, and Jinwoo Shin. Csi: Novelty detection via contrastive learning on distributionally shifted instances. *Advances in Neural Information Processing Systems*, 33, 2020. 5, 6

[42] 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 conference on computer vision and pattern recognition*, pages 8769–8778, 2018. 7

[43] Sagar Vaze, Kai Han, Andrea Vedaldi, and Andrew Zisserman. Open-set recognition: A good closed-set classifier is all you need. In *ICLR*, 2022. 2, 3, 4, 5, 6, 7, 8

[44] Haoqi Wang, Zhizhong Li, Litong Feng, and Wayne Zhang. Vim: Out-of-distribution with virtual-logit matching. In *CVPR*, 2022. 1, 4, 5, 6, 7, 12

[45] Hongxin Wei, Renchunzi Xie, Hao Cheng, Lei Feng, Bo An, and Yixuan Li. Mitigating neural network overconfidence with logit normalization. In *ICML*, 2022. 1, 2, 3, 4, 5

[46] 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 *Thirty-sixth Conference on Neural Information Processing Systems Datasets and Benchmarks Track*, 2022. 2, 4, 5, 6, 7

[47] Jingkang Yang, Kaiyang Zhou, Yixuan Li, and Ziwei Liu. Generalized out-of-distribution detection: A survey. *arXiv preprint arXiv:2110.11334*, 2021. 2

[48] Dong Yin, Raphael Gontijo Lopes, Jon Shlens, Ekin Dogus Cubuk, and Justin Gilmer. A fourier perspective on model robustness in computer vision. *Advances in Neural Information Processing Systems*, 32, 2019. 8

[49] Sangdoo Yun, Dongyoon Han, Seong Joon Oh, Sanghyuk Chun, Junsuk Choe, and Youngjoon Yoo. Cutmix: Regularization strategy to train strong classifiers with localizable features. In *CVPR*, 2019. 1, 4, 5

[50] 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)*, pages 5531–5540, January 2023. 1

[51] Bolei Zhou, Agata Lapedriza, Aditya Khosla, Aude Oliva, and Antonio Torralba. Places: A 10 million image database for scene recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2017. 5