Mixture Outlier Exposure: Towards Out-of-Distribution Detection in Fine-grained Environments

Jingyang Zhang[†], Nathan Inkawhich[*], Randolph Linderman[†], Yiran Chen[†], Hai Li[†]
[†]Duke University, [*]Air Force Research Laboratory
jingyang.zhang@duke.edu

Abstract

Many real-world scenarios in which DNN-based recognition systems are deployed have inherently fine-grained attributes (e.g., bird-species recognition, medical image classification). In addition to achieving reliable accuracy, a critical subtask for these models is to detect Out-of-distribution (OOD) inputs. Given the nature of the deployment environment, one may expect such OOD inputs to also be fine-grained w.r.t. the known classes (e.g., a novel bird species), which are thus extremely difficult to identify. Unfortunately, OOD detection in fine-grained scenarios remains largely underexplored. In this work, we aim to fill this gap by first carefully constructing four large-scale fine-grained test environments, in which existing methods are shown to have difficulties. Particularly, we find that even explicitly incorporating a diverse set of auxiliary outlier data during training does not provide sufficient coverage over the broad region where fine-grained OOD samples locate. We then propose Mixture Outlier Exposure (MixOE), which mixes ID data and training outliers to expand the coverage of different OOD granularities, and trains the model such that the prediction confidence linearly decays as the input transitions from ID to OOD. Extensive experiments and analyses demonstrate the effectiveness of MixOE for building up OOD detector in fine-grained environments. The code is available at `https://github.com/zjysteven/MixOE`.

1. Introduction

Real-world scenarios in which DNN recognition systems are deployed are often fine-grained in nature, wherein the data coming from such environments share a high level of semantic/visual similarity. Examples include fine-grained visual classification [49, 13, 46], medical image classification [23, 7], and remote sensing applications [50, 45, 44]. In addition to achieving accurate classification, a critical problem for DNN models is to identify out-of-distribution (OOD) samples during inference time which do not belong to one of the DNN's known classes. Such detection is crucial for building safe and reliable intelligent systems that operate in the open world. However, we posit that OOD detection is particularly challenging in fine-grained scenarios because one may expect the OOD inputs to be highly granular w.r.t. the in-distribution (ID) data (*e.g.*, novel bird species to a bird classifier), given the nature of the deployment environment. Such fine-grained OOD samples can make detection much difficult because they may use very similar feature sets to the ID data [16].

Unfortunately, despite being inherently motivated by many real-world scenarios, *OOD detection in fine-grained environments remains largely underexplored/underconsidered in current research.* This in part has to do with (overly) simplistic [1, 37] and coarse-grained benchmarks that are currently being used to evaluate OOD detection methods [25, 22, 36, 26, 14, 43, 38] (*e.g.*, CIFAR-10/100 v.s. SVHN/LSUN). An illustrative comparison between OOD detection in fine- and coarse-grained environments is shown in Fig. 1. A few recent works [31, 1, 39] did consider fine-grained settings but did *not* carefully/thoroughly investigate why and how they are difficult. Besides, these works either operated at a rather limited scale, *i.e.*, very few ID classes were considered [1, 39], or made unrealistic practices such as training on a *labeled* outlier dataset that overlaps with testing OOD data distribution [31].

In this work, *our goal is to fill the gap and present (to our knowledge) the first study that specifically targets OOD detection in fine-grained environments.* We start by building four large-scale, fine-grained test environments to approximate real-world scenarios (Sec. 3). We find that several state-of-the-art OOD detection methods struggle to detect fine-grained novelties, which highlights the challenging nature of OOD detection in fine-grained settings. Then, through analysis we identify that fine-grained OOD samples span a much broader region and are closer to the ID clusters in the DNN's feature space (Sec. 4.1). In addition, we find that including a large/diverse set of outlier data dur-

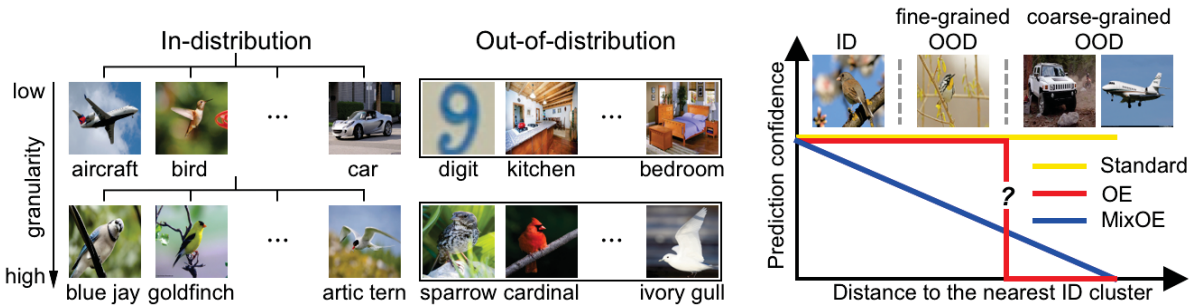

Figure 1. **Left:** A comparison of OOD detection in coarse- and fine-grained environments. Intuitively, fine-grained detection is much more challenging. **Right:** A *conceptual* illustration of MixOE. A Standard model with no OOD considerations tends to be over-confident on OOD samples. OE is able to calibrate the prediction confidence on coarse-grained OOD, but the outputs on fine-grained OOD are not explicitly controlled (marked by "?"). MixOE aims for a smooth decay of the confidence as the inputs transition from ID to OOD, and thus enables detection of both coarse/fine-grained OOD.

ing training [12, 2, 26] does *not* help cover the area where fine-grained OOD data locate.

Finally, we design a novel training algorithm named *Mixture Outlier Exposure* (MixOE) to address the observed issue in fine-grained OOD detection (Sec. 4.2). Specifically, we propose to perform *mixing operations* (in this work we adopt Mixup [48] or CutMix [47]) between ID data and outlier data to get "virtual" outlier samples which can be both near to and far away from the ID clusters. The model is then trained such that the prediction confidence linearly decays as the input transitions from ID to OOD (see Fig. 1 for illustration). As such, MixOE induces regularization over a larger OOD region and has clear implications for both coarse- and fine-grained OOD samples. Experimental results on the four test benches show that a simple fine-tuning with MixOE can lead to consistently higher or competitive detection rates against both coarse- and fine-grained OOD data (Sec. 5). We also conduct careful ablation study to further understand why and how MixOE works. Our contributions are summarized as follows:

- We construct four large-scale test environments for fine-grained OOD detection, which can be easily generated from existing public datasets and facilitate future studies;
- We propose MixOE, a novel OOD detection methodology that has effect across a spectrum of OOD granularities;
- We show that MixOE leads to notable improvements on all four benchmarks, in particular against fine-grained OOD where few current methods have any impacts.

2. Related work

OOD detection approaches. Many popular works in OOD detection research use a pre-trained DNN classifier as a base model, and design an OOD scoring mechanism that leverages some signal from this model. Several methods utilize the output space of the classifier, *e.g.*, MSP [12], ODIN [25], and Energy [26], while other works such as Mahalanobis detector [22] and Gram Matrices [36] focus on the intermediate feature space of DNNs. Recent works also start to explore the potential of gradient information [15].

Another line of research modifies the base DNN's training phase to enable better detection. Lee *et al.* [21] proposed to synthesize OOD samples with a GAN [32] and force the classifier to be less confident on the generated OOD data. Hendrycks *et al.* [12] later showed that a diverse and realistic outlier distribution is more proficient than synthetic samples at encouraging low confidence predictions on unseen OOD data. More recently there comes the idea of VOS [6], which does not use real data but generates virtual outliers by sampling from class-conditional Gaussian distribution. However, the assumption that ID data follow class-conditional Gaussian may not hold especially when the number of classes is large, and VOS has shown to even have trouble scaling to CIFAR-100 [18]. To conclude, incorporating auxiliary outlier data in training is still one of the most effective methods to date [12, 2, 26].

Our work is most closely related to [12] in that we too use *unlabeled* auxiliary outlier data. However, our method uniquely formulates the learning procedure that has explicit consideration of operating in a fine-grained setting where highly granular OOD inputs are expected. We also remark that the above works all consider relatively coarse-grained settings in their experiments.

Mixing operations for OOD detection. Our work also relates to, but differs significantly from a few works that utilized mixing operations in the context of OOD detection. The work of [40] and [4] both directly evaluated Mixup's effect on OOD detection. In this work, instead of plainly applying Mixup or CutMix as a regular ID training strategy, we leverage them to construct the virtual outlier distribu-

tion to expand the coverage over OOD region. Ravikumar *et al.* [33] proposed to apply Mixup either between ID samples or between the training outliers. In contrast, in our framework the mixing operations are performed between ID and outlier data, which has explicit implication in characterizing the transition from ID to OOD region. In Sec. 5.3, we will demonstrate the superiority of the proposed MixOE over these methods through careful ablation study.

OOD detection in fine-grained settings. As aforementioned, there are a few works that lightly considered OOD detection in fine-grained environments, yet they all have significant limitations. In the work of [31], a *labeled* outlier dataset was used for training, which we believe is a prohibitive assumption in reality. Even more concerning, in its experiments the training outlier dataset (ImageNet [5]) overlapped with the testing OOD data (CUB [42] and Stanford Dogs [19]), which makes the detection (arguably) trivial. The work of [1] and [39] are limited in that the scale of the fine-grained detection problems in their experiments is rather small, *i.e.*, only tens of or even fewer ID classes were considered, whereas we operate at a much larger scale with hundreds of ID classes. Later, we will also show that the method studied in [1] is not yet effective for detecting fine-grained OOD examples.

We also notice some recent works that purposefully consider cases where the test OOD samples are semantically related to the ID classes [43, 17]. The critical difference here is that they did *not* consider the ID classes to be granular. We argue that without such ID granularity, it may be debatable whether a model *should* be asked to detect granular OOD samples! For example, consider an experiment in [43] that regards "leopard" (in CIFAR-100) as a granular OOD sample for a CIFAR-10 model (which contains a "cat" class). This situation begs the question: given a model trained on cats, is it more desirable to generalize to the notion of a leopard[1], or to identify leopards as OOD? In our work, since we consider ID classes that are themselves highly granular (*e.g.*, unique bird species), such ambiguity is avoided and the fine-grained novelties (*e.g.*, novel bird species) should be considered OOD.

3. Challenges of OOD detection in fine-grained settings

This section describes a detailed study on OOD detection in fine-grained settings and serves to further motivate our goal of improving detection explicitly in fine-grained scenarios. In Sec. 3.1 we describe the construction of the four fine-grained test environments, which is necessary because those presented in prior works are limited in scale. In Sec. 3.2, through initial evaluations we show that fine-grained OOD detection presents distinct challenges for existing methods.

Table 1. Comparison of constructed fine-grained OOD settings. For example, in our third test environment there are 150 ID and 46 (fine-grained) OOD categories. We consider a larger scale than previous works.

	# ID classes	# OOD classes
[31]	[100, 60]	[100, 60]
[1]	[11, 9, 8, 7, 7]	[1, 1, 1, 1, 1]
[39]	[46, 20]	[47, 5]
Ours	[200, 150, 150, 90]	[55, 50, 46, 12]

3.1. Test environments

The test environments are curated from four public fine-grained visual classification (FGVC) datasets, namely FGVC-Aircraft [28], Stanford Cars [20], Butterfly [3], and North American Birds [41]. We refer to them as *Aircraft*, *Car*, *Butterfly*, and *Bird*, respectively. For each dataset, we create ID/OOD splits using a holdout class method, *i.e.*, we keep some of the categories as ID and the rest are held out from the training set and considered OOD at test time. Note, to avoid implicit bias that might exist in each single split [1], we randomly produce three ID/OOD splits for each dataset with equal counts of ID/OOD categories. See Appendix A for more details including exact splits and number of train/test images.

In Tab. 1 we present a comparison between our constructed environments and previous ones. By considering 100+ ID classes we are operating at a larger scale which better represents a wide variety of complex real-world tasks and avoids putting any restrictive assumptions about the complexity of the ID classification task that would make our findings less scalable. We also leave a reasonable number of classes as OOD to reflect the diversity of the open world.

Besides fine-grained novel inputs, a reliable detector should also be able to identify coarse-grained OOD data. Here, for each dataset, we take the images from the other datasets as coarse-grained OOD samples (*e.g.*, when *Bird* is ID, *Butterfly*, *Car*, and *Aircraft* will be considered OOD). The detection performance is then evaluated against both fine- and coarse-grained novelties.

3.2. Evaluating existing methods

Setup. We now evaluate six state-of-the-art detectors in the constructed fine-grained environments, including three post-training scorers[2] (MSP [11], ODIN [25], and Energy [26]) and three that incorporate training-time regularizations using auxiliary outlier data (OE [12], OE with hard

[1]This is actually the desired behaviour in the problem of subpopulation shift robustness [35].

[2]We also evaluated Mahalanobis detector [22] but simply got NaN error, which aligns with [10]'s finding that it has scalability issue.

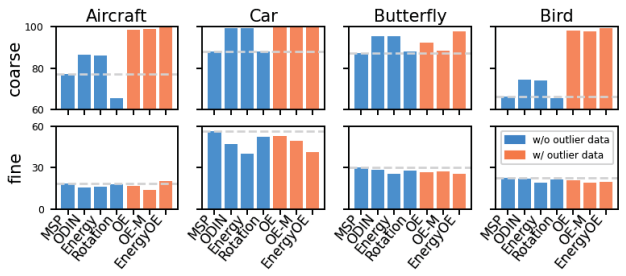

Figure 2. TNR95 of existing methods against coarse-grained (first row) and fine-grained OOD data (second row). The gray dashed line is the baseline performance (MSP). Fine-grained novelties are significantly harder to detect in all datasets for all methods. Also note how the methods that utilize outlier data (coral ones) help with coarse-grained OOD but barely improve fine-grained detection rates.

example mining (OE-M) [2], and EnergyOE [26]). In addition, we consider the method Rotation [1] which was studied in a related work that mentioned fine-grained detection. We leave implementation details of the detection methods to Sec. 5.1 and Appendix B.

To measure detector performance, we use two common metrics [22, 21, 14]: true negative rate at a 95% true positive rate (TNR95) and area under the receiver operating characteristic curve (AUROC). While AUROC is a holistic measurement obtained by varying the threshold, TNR95 indicates what portion of OOD samples could be detected when the recall of ID data is 95%. TNR95 is known to better separate different detectors since achieving a high TNR95 is much harder than achieving a high AUROC [14].

Observations. Fig. 2 shows the TNR95 results on one of the three splits in each dataset (other splits present similar patterns; see Appendix E Fig. 7). Also see the full result table including AUROC statistics in Appendix E Tab. 6. From these results we make two important observations. First, fine-grained OOD samples are significantly more difficult to detect than coarse-grained ones. Specifically, while most methods can achieve more than around 80% TNR95 when detecting coarse-grained OOD samples (Fig. 2 first row), the TNR95 drops to below 30% for all methods on 3 of the 4 datasets when facing fine-grained OOD (Fig. 2 second row). This observation is in line with recent findings that detection becomes more challenging when OOD data are semantically similar to ID classes [14, 43, 17].

Our second observation is more surprising: on the hold-out fine-grained data, even the methods that explicitly include outlier data during the training (OE, OE-M, and EnergyOE) do not reliably outperform MSP, which trains the model using ID data only. This finding directly contrasts the results on coarse-grained data, where OE/OE-M/EnergyOE consistently lead to improvements (compared with MSP). In all, these trends clearly demonstrate that OOD detection in fine-grained settings with highly granular OOD inputs is particularly challenging for existing detectors.

4. Methodology

4.1. Motivating analysis

An immediate question that arises from our initial results is *why using auxiliary outlier data to explicitly regularize the model helps with detecting coarse-grained OOD samples but not fine-grained ones?* In fact, the training outlier data we use do *not* characterize/model the test coarse- or fine-grained OOD at all (see Fig. 3; we manually ensure this via filtering as discussed in Sec. 5.1). Next, we conduct analysis to provide an explanation to this question, which also directly motivates our methodology.

The idea of our analysis is to reveal the relationship between the training outliers and test OOD samples by projecting them into the DNN's feature space. Specifically, we forward pass the train/test OOD data into a "standard" pre-trained model, take the outputs of the penultimate layer as their features, and visualize them in a 2D space using the technique in [30] (see Appendix C for details). Critically, the visualization process is uniquely formed by ID data and thus remains *agnostic* to OOD data. The standard model here is trained using cross-entropy loss on ID data only. Since OE-based methods typically fine-tune the standard model using the outlier data [12, 26, 2], by visualizing in the standard model's feature space we can anticipate/explain how the training outliers may help regularize the model.

Fig. 3 shows the results on one of the test benches (see Appendix E Fig. 12-15 for more), from which we make two key observations. First, as shown in Fig. 3 (b), coarse-grained OOD data locate in a rather compact region, with a small portion of samples intersecting with one of the ID clusters and others being relatively far away from the ID region. According to Fig. 3 (c), however, fine-grained OOD samples span a much broader area, with many of them being very close to or even within the ID clusters, due to their semantic similarity to the ID images. This observation reinforces the TNR95 detection results in Fig. 2, where fine-grained novelties are significantly more difficult to detect than coarse-grained ones.

Second, comparing Fig. 3 (d) and (b), the auxiliary outlier data "enclose" the coarse OOD data region. As a result, although the outliers do not have any concepts related to the test coarse OOD samples (*i.e.*, no cars/butterflies/aircrafts in Fig. 3 (d) second row), their regularization still generalize to the test coarse OOD inputs. On the other hand, clearly the training outliers fail to cover the larger area where many fine-grained OOD data locate, which explains why methods utilizing the outlier data have limited effect in detecting fine-grained novelties.

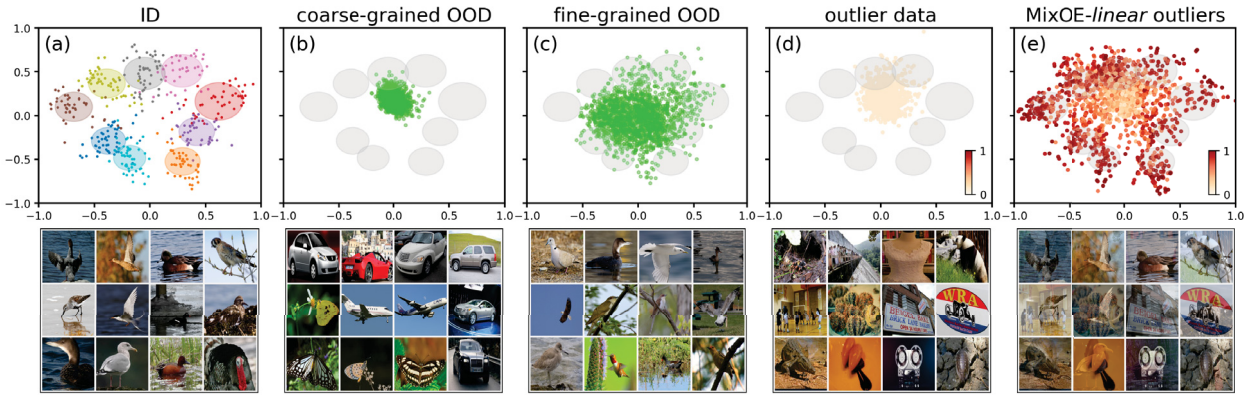

Figure 3. Visualization of the data samples (second row) and their representations in the DNN's feature space (first row). The color lightness in (d)/(e) indicates the prediction confidence encoded in the soft target of each corresponding outlier sample. Note, (b) and (c) are test OOD samples and are never seen during the training. The empirical training outliers (d) enclose the region where coarse OOD data (b) locate but does not cover the much broader area the fine OOD samples (c) span. MixOE mixes the ID (a) and training outliers (d) to induce larger coverage which accounts for both coarse- and fine-grained novelties. Moreover, the soft targets of the mixed data will calibrate the model's prediction confidence to smoothly decay from ID to OOD.

4.2. Mixture Outlier Exposure

To explicitly regularize the model's behaviour in a broader region to improve detection against both coarse- and fine-grained OOD data, we would like to create a "virtual" outlier distribution $\mathcal{D}_{\text{out}}^{\text{virtual}}$ which can provide a more comprehensive coverage over the OOD region than the empirical outlier distribution $\mathcal{D}_{\text{out}}$.

Generating mixed outliers. Specifically, given an ID sample $(x_{\text{in}}, y_{\text{in}}) \sim \mathcal{D}_{\text{in}}$ and an outlier sample $x_{\text{out}} \sim \mathcal{D}_{\text{out}}$, we propose to perform *mixing operations* to generate the virtual outliers:

$$\tilde{x} = \text{mix}(x_{\text{in}}, x_{\text{out}}, \lambda), \tag{1}$$

where $\lambda \in [0, 1]$ is a coefficient controlling the contribution of each sample to the mixed one (*i.e.*, λ and $1 - \lambda$ are the weight for x_{in} and x_{out}, respectively). The intuition here is directly based on our observations in Fig. 3: If we "interpolate" the ID samples (Fig. 3 (a)) and outlier data (Fig. 3 (d)), the resulting samples are likely to span a larger area and cover the region where fine-grained OOD data locate (Fig. 3 (c)).

We find that simple pixel-level operations, *e.g.*, linear interpolation [48] and cut-paste operation [47], can already result in samples that cover the fine-grained OOD region. Thus in this work we instantiate Eqn. 1 with these operations and denote them as *linear* mixing and *cut* mixing, respectively. However, we remark that this operation may be as simple or as complex as desired; it is not constrained to [48, 47] and allows possible extension in the future.

To demonstrate the effect of mixed samples $\tilde{x}$, we again visualize them in the DNN model's feature space in Fig. 3 (e) (here we use *linear* mixing as an example; see Appendix E Fig. 12-15 for more). Importantly, unlike the empirical outlier distribution $\mathcal{D}_{\text{out}}$ in Fig. 3 (d), our virtual outlier samples can span a larger area (being *both* near to and far away from the ID clusters) by varying the coefficient λ. As a result, we anticipate that when training with the samples from $\mathcal{D}_{\text{out}}^{\text{virtual}}$, the model's behaviour on both fine- and coarse-grained OOD data can be regularized.

Assigning training targets. Now that we have generated mixed outliers for training, the next important step is to decide the corresponding training targets. *Our key insight here is to regularize the model such that its prediction confidence can smoothly decay as the input transitions from ID to OOD.* To this end, we assign soft target $\tilde{y}$ corresponding to the mixed sample $\tilde{x}$ as follows:

$$\tilde{y} = \lambda y_{\text{in}} + (1 - \lambda)\mathcal{U}, \tag{2}$$

where y_{in} is the one-hot label of the ID sample x_{in}, and $\mathcal{U}$ represents the uniform distribution over the ID categories. The prediction confidence (*i.e.*, maximum softmax probability [11]) encoded in $\tilde{y}$ is $\tilde{y}_{\text{pred. conf.}} = \lambda + (1 - \lambda)\frac{1}{K}$, where K is the number of ID categories. Concretely, when the mixed sample is OOD (*i.e.*, $\lambda = 0$), we force the model to be least confident/maximally uncertain on that sample (*i.e.*, $\tilde{y}_{\text{pred. conf.}} = \frac{1}{K}$); when the mixed sample is ID (*i.e.*, $\lambda = 1$), the model is trained to make a confident prediction (*i.e.*, $\tilde{y}_{\text{pred. conf.}} = 1$). The confidence of the intermediate mixed samples (*i.e.*, $\lambda \in (0, 1)$) is smoothly modulated by λ. A visualization of such effect can be seen in Fig. 3 (e), where darker points correspond to a higher confidence that is encoded in their soft targets $\tilde{y}$. Unlike previous methods, we remark that this formulation uniquely enables the model to have better calibration over a wider range of confidence levels. As a result, our method can have effects across a spectrum of OOD granularities, which is crucial for detectors in fine-grained environments as the OOD samples could be highly granular.

Training objective. The above two technical insights make up our training algorithm, Mixture Outlier Exposure (MixOE), whose objective is formulated as

$$\mathbb{E}_{(x,y)\sim\mathcal{D}_{\text{in}}}\big[\mathcal{L}(f(x),y)\big]+\beta\mathbb{E}_{(\tilde{x},\tilde{y})\sim\mathcal{D}_{\text{out}}^{\text{virtual}}}\big[\mathcal{L}(f(\tilde{x}),\tilde{y})\big]. \tag{3}$$

Here, $\mathcal{L}(f(x),y)$ is the cross-entropy loss between the DNN's predicted distribution $f(x)$ and the ground truth distribution y, and β is a weighting term. During training, at each iteration the λ in Eqn. 1 and Eqn. 2 is sampled from a Beta distribution $\text{Beta}(\alpha,\alpha)$ for an introduced hyperparameter α. Both α and β can be determined using validation data (details are in Sec. 5.1). We also remark here that the outlier dataset used to construct $\mathcal{D}_{\text{out}}^{\text{virtual}}$ is *unlabeled* and does *not* need to contain any data related to the test OOD data distribution. After training, the detection will be performed by thresholding the prediction confidence since MixOE explicitly calibrates the confidence on outliers during the training.

Relationship with prior methods. The training objective of the vanilla OE [12] is $\mathbb{E}_{(x,y)\sim\mathcal{D}_{\text{in}}}\big[\mathcal{L}(f(x),y)\big]+\beta\mathbb{E}_{x_{\text{out}}\sim\mathcal{D}_{\text{out}}}\big[\mathcal{L}(f(x_{\text{out}}),\mathcal{U})\big]$, which encourages the model's output to resemble uniform distribution $\mathcal{U}$ on the empirical outliers from $\mathcal{D}_{\text{out}}$. What separates MixOE from OE is the novel idea of using generated virtual outliers to expand the regularization over a broader region and controlling how the confidence decay from ID to OOD. Also, note that MixOE can degenerate to vanilla OE if λ is fixed to 0 during the training. The other two methods, OE-M [2] and EnergyOE [26], share the same idea as OE except that [2] uses the "hardest" outliers from $\mathcal{D}_{\text{out}}$ and [26] adopts a different scoring function for detection. They still uses only the empirical outliers and has no control over the confidence decay. Thus we believe they share similar shortcomings to OE, which we will confirm with experimental results shortly.

5. Experiments

5.1. Setup

Baselines. Same as in Sec. 3.2, we consider a total of seven baseline methods, including six state-of-the-art methods [11, 25, 26, 12, 2] and one that was studied in a previous fine-grained setting [1]. We also re-emphasize that there are very few methods that have considerations for operating in fine-grained environments as this is an underexplored topic.

Training details. For the post-training scoring methods, we train ResNet-50 models [9] in the standard fashion, *i.e.*, by minimizing the cross-entropy loss over the ID-only training dataset. Specifically, we train the model for 90 epochs using SGD with the batch size being 32. Following common practices in fine-grained classification research [49, 13, 46], the model is initialized with ImageNet pre-trained weights. The initial learning rate is 0.001 and is decayed by cosine learning rate schedule [27]. For Rotation [1], we train the model using its objective with the same setup as the standard training.

For methods that utilize auxiliary outlier data (OE, OE-M, EnergyOE, and MixOE), we only fine-tune the trained standard model with the corresponding objective for 10 epochs, following [12, 26]. Therefore, MixOE and other OE-based methods only induce marginal computation overhead. The fine-tuning also adopts cosine schedule with the initial learning rate being 0.001. The batch size of ID data is still 32. For OE/OE-M/EnergyOE, as suggested in their papers, we set the batch size of outlier data to be twice as the ID batch size, which is 64. In the case of MixOE, we keep outlier batch size same as the ID batch size. As a result, MixOE actually uses only half of the outliers used by other methods.

The auxiliary outlier set. The training outlier set $\mathcal{D}_{\text{out}}$ we consider is WebVision 1.0 [24], which contains natural images crawled from Flickr and Google by querying with the 1,000 categories of ImageNet. We believe this dataset represents a realistic and practical construction of $\mathcal{D}_{\text{out}}$ for many ID tasks in the natural imagery domain.

Importantly, to avoid arguments of "cheating", *we filter out images that are relevant to the considered OOD tasks from the outlier set based on WordNet ID* [29]. Specifically, a total of 491K images related to aircraft/car/butterfly/bird are removed. Thus, the training outliers reveal no information about the test coarse- or fine-grained OOD data (see Fig. 3 for visualization). After the filtering, there are 1948K images left in the outlier set. However, since MixOE just fine-tunes the model for 10 epochs, at most 70K images are actually used during the training.

Hyperparameter tuning. We take great care to ensure that the hyperparameter tuning is fair. Concretely, we randomly holdout a portion of samples from the ID and outlier training set to serve as ID/OOD validation data. Critically, note that the OOD validation data reveals no information about the test-time OOD distribution since we already filter out all relevant images from $\mathcal{D}_{\text{out}}$.

With the selected ID and OOD validation samples, we tune the hyperparameters such that the OOD detection performance is maximized and the ID classification accuracy is minimally affected. To test the method robustness, in each of the four environments we only tune the hyperparameter once for each approach using a single split; the determined hyperparameter is then applied to all splits from the same dataset. In Appendix B we present a detailed list of the candidate hyperparameter values and the final deteremined values we use for each method in our experiments.

Evaluation. The evaluation procedure follows the one described in Sec. 3.1 and Sec. 3.2. For each dataset, we consider the holdout classes as fine-grained OOD data and the samples from other datasets as coarse-grained OOD data.

Table 2. Detection performance in terms of TNR95 statistics. The number before and after the slash is for coarse-grained and fine-grained OOD samples, respectively. Avg. diff. is the average difference (across three splits) relative to MSP. Clearly, MixOE consistently leads to notable improvements against both coarse- and fine-grained novelties, while other methods degrade the fine-grained OOD detection performance and thus do not qualify for a reliable detector in fine-grained environments.

$\mathcal{D}_{\text{in}}$	Method	Split 1	Split 2	Split 3	Avg. diff.
Aircraft	MSP [11]	75.0 / 29.9	61.6 / 15.9	77.1 / 18.5	- / -
	ODIN [25]	87.5 / 30.2	73.2 / 15.3	86.5 / 15.8	+11.2 / –1.0
	Energy [26]	88.5 / 30.1	74.4 / 14.6	86.2 / 16.3	+11.8 / –1.1
	Rotation [1]	65.5 / 31.4	55.0 / 15.9	65.5 / 17.6	–9.2 / +0.2
	OE [12]	99.3 / 27.8	98.5 / 16.0	98.7 / 16.5	+27.6 / –1.3
	OE-M [2]	99.6 / 25.0	98.5 / 16.0	98.9 / 14.0	+27.8 / –3.1
	EnergyOE [26]	99.8 / 30.3	99.7 / 17.0	99.7 / 19.9	+28.5 / +1.0
	MixOE-*linear*	93.2 / 41.4	88.4 / 24.6	92.1 / 16.5	+20.0 / +6.1
	MixOE-*cut*	99.0 / 39.8	99.4 / 23.7	99.4 / 24.9	+28.0 / +8.0
Car	MSP [11]	95.5 / 58.5	88.0 / 56.3	78.8 / 53.5	- / -
	ODIN [25]	99.6 / 55.6	99.1 / 47.0	97.8 / 49.0	+11.4 / –5.6
	Energy [26]	99.7 / 49.1	99.4 / 39.7	99.1 / 42.6	+12.0 / –12.3
	Rotation [1]	97.7 / 58.9	88.1 / 52.4	81.3 / 50.4	+1.6 / –2.2
	OE [12]	99.9 / 53.2	100.0 / 53.0	99.9 / 51.2	+12.5 / –3.6
	OE-M [2]	99.9 / 53.6	100.0 / 49.4	100.0 / 50.6	+12.5 / –4.9
	EnergyOE [26]	100.0 / 52.6	100.0 / 41.0	100.0 / 44.9	+12.6 / –9.9
	MixOE-*linear*	99.6 / 65.9	99.7 / 62.9	99.5 / 60.1	+12.2 / +6.9
	MixOE-*cut*	99.9 / 70.3	100.0 / 69.8	99.9 / 66.5	+12.5 / +12.8
Butterfly	MSP [11]	87.1 / 29.9	89.9 / 31.8	88.4 / 36.6	- / -
	ODIN [25]	95.2 / 28.2	95.5 / 32.5	95.6 / 38.7	+7.0 / +0.4
	Energy [26]	95.3 / 25.5	95.2 / 30.2	95.6 / 36.1	+6.9 / –2.2
	Rotation [1]	87.9 / 27.6	88.5 / 31.2	86.2 / 37.0	–0.9 / –0.8
	OE [12]	92.2 / 26.5	93.7 / 32.1	94.3 / 34.3	+4.9 / –1.8
	OE-M [2]	99.9 / 53.6	100.0 / 49.4	100.0 / 50.6	+12.5 / –4.9
	EnergyOE [26]	97.8 / 25.1	96.9 / 30.5	98.2 / 37.2	+9.2 / –1.8
	MixOE-*linear*	95.3 / 32.6	93.9 / 37.9	95.5 / 45.0	+6.4 / +5.7
	MixOE-*cut*	94.9 / 35.8	94.1 / 38.8	92.7 / 46.0	+5.4 / +7.4
Bird	MSP [11]	72.3 / 22.6	67.4 / 22.3	66.4 / 22.3	- / -
	ODIN [25]	80.9 / 22.7	77.2 / 21.5	74.3 / 21.9	+8.8 / –0.4
	Energy [26]	80.8 / 20.3	76.5 / 18.4	73.9 / 18.8	+8.4 / –3.2
	Rotation [1]	71.3 / 23.6	64.0 / 24.0	65.4 / 21.5	–1.8 / +0.6
	OE [12]	98.2 / 20.6	97.9 / 22.9	97.9 / 20.7	+29.3 / –1.0
	OE-M [2]	98.7 / 19.8	98.7 / 21.4	97.7 / 19.2	+29.7 / –2.3
	EnergyOE [26]	98.6 / 19.4	99.0 / 18.4	99.3 / 19.5	+30.3 / –3.3
	MixOE-*linear*	88.6 / 24.9	83.9 / 26.7	86.3 / 28.6	+17.6 / +4.3
	MixOE-*cut*	91.0 / 27.7	91.8 / 24.6	92.9 / 27.7	+23.2 / +4.3

Following [14, 22, 21], we consider ID as positive and OOD as negative, and use TNR95 and AUROC as the metrics.

5.2. Results

Detection performance. Tab. 2 shows the TNR95 results across the four test benches. The AUROC statistics yield similar patterns to TNR95 and are left in the expanded Tab. 6 in Appendix E.

Our first observation is that MixOE consistently achieves the best detection performance against fine-grained OOD samples. Specifically, averaged across the three splits on the [*Aircraft*, *Car*, *Butterfly*, *Bird*] tasks, MixOE-*linear* and MixOE-*cut* improve the TNR95 over MSP by [+6.1%, +6.9%, +5.7%, +4.3%] and [+8.0%, +12.8%, +7.4%, +4.3%], respectively. In comparison, vanilla OE's relative improvement over MSP is [–1.3%, –3.6%, –1.8%, –1.0%]. Similar to OE, the other baseline methods also result in performance degradation in the face of fine-grained OOD on many (if not all) datasets. We also remark that MixOE-*linear* and MixOE-*cut* both lead to significant improvements, demonstrating that the validity of the idea behind MixOE is independent of the specific mixing operations being used.

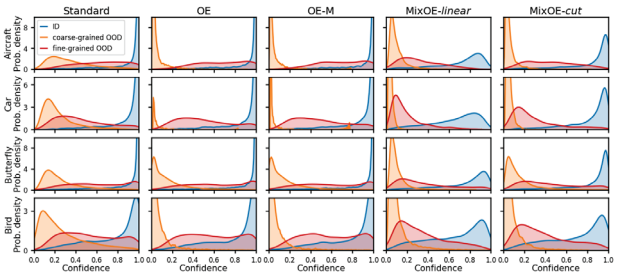

Figure 4. Comparison of the prediction confidence's distribution between methods. MixOE leads to clearer separation between the confidence of ID and OOD samples (especially fine-grained ones) and thus enables better detection.

Our second observation is that MixOE can perform on par with state-of-the-art methods in detecting coarse-grained OOD samples. On the four tasks MixOE-*linear* and MixOE-*cut* improve the TNR95 over MSP by [+20.0%, +12.2%, +6.4%, +17.6%] and [+28.0%, +12.5%, +5.4%, +23.2%], respectively. OE, whose effect is exclusively towards coarse-grained OOD data, leads to improvements of [+27.6%, +12.5%, +4.9%, +29.3%].

Finally, we note that MixOE is the only method that consistently improves upon MSP against both fine- and coarse-grained novelties across all the datasets. This unique capability of MixOE to remain effective across a spectrum of OOD granularities is critical for systems that operate in fine-grained environments, as the novel inputs during inference can be either coarse or fine. Overall, these evaluation results display the effectiveness of MixOE for building up reliable OOD detectors in real-world fine-grained settings.

Prediction confidences. To more closely understand how MixOE improves the OOD detection, we monitor the confidence distributions of the models on ID/OOD samples as MixOE functions by calibrating the predication confidence. Fig. 4 shows the probability density plots of prediction confidence on Standard, OE, OE-M, and MixOE models on one of the splits from each dataset (see Appendix E Fig. 8-11 for more). From Fig. 4 we can clearly identify that, regardless of the mixing operation, the MixOE models consistently produce lower confidence predictions on fine-grained OOD samples, making them more distinguishable from ID inputs. This observation confirms that the virtual outlier data and their corresponding soft targets introduced in MixOE indeed help regularize the model's outputs on fine-grained OOD samples.

Table 3. Accuracy comparison of the training algorithms. The numbers in the parenthesis are the differences relative to the accuracy of standard training. Avg. diff. shows the improvements averaged across the four datasets. Unlike other methods that trade-off accuracy for detection, MixOE acutally improves the accuracy.

	Aircraft	Car	Butterfly	Bird	Avg. diff.
Rotation [1]	88.5 (–1.1)	91.3 (–0.6)	88.8 (–0.1)	82.0 (–0.1)	–0.5
OE [12]	89.2 (–0.5)	91.6 (–0.2)	88.1 (–0.9)	82.4 (+0.3)	–0.3
OE-M [2]	89.3 (–0.4)	91.1 (–0.8)	88.2 (–0.8)	82.7 (+0.6)	–0.4
EnergyOE [26]	89.3 (–0.3)	91.8 (–0.0)	88.8 (–0.2)	82.3 (+0.2)	–0.1
MixOE-*linear*	90.5 (+0.8)	92.9 (+1.1)	89.3 (+0.3)	83.4 (+1.3)	+0.9
MixOE-*cut*	90.1 (+0.5)	92.9 (+1.1)	90.1 (+1.2)	83.5 (+1.4)	+1.1

ID classification accuracy. Lastly, we examine how the training methods can affect the ID classification accuracy because we do not intend to tradeoff accuracy for detection performance. In Tab. 3 we show the accuracy on each dataset averaged across the three splits (see Appendix E Tab. 7 for unaveraged results). Interestingly, unlike other training strategies, MixOE-*linear* and MixOE-*cut* can improve the accuracy by 0.9% and 1.1% on average across the four environments, respectively. Our hypothesis here is that since the fine-grained datasets often have relatively small number of training samples (*e.g.*, tens of images per class), some of the training "outliers" generated by MixOE that are close to the ID clusters actually serve as augmented data and thus contribute to the ID accuracy.

5.3. Ablation study

MixOE v.s. Mix. Recall that the core idea of MixOE is to mix ID and outlier data for DNN training. In Sec. 5.2 we have shown that this concept is more beneficial than using the outlier data alone without mixing (vanilla OE). Here we ablate MixOE along another direction: We contrast MixOE with vanilla Mix training [48, 47] which does not use the auxiliary outlier data, *i.e.*, mixing only ID data. Concretely, for Mix training the generation of virtual outlier samples in Eqn. 3 is changed to $(\tilde{x}, \tilde{y}) = (\text{mix}(x_1, x_2, \lambda), \lambda y_1 + (1 - \lambda) y_2)$, where $(x_1, y_1), (x_2, y_2) \sim \mathcal{D}_{\text{in}}$. The hyperparameter tuning procedure and training setup is the same as those for MixOE.

The comparison is presented in Fig. 5, where we show the methods' average improvements in TNR95 relative to the baseline MSP across the three splits on each dataset. We find that Mix training is able to provide performance gains in detecting fine-grained OOD data, but the gains are smaller than those brought by MixOE; meanwhile, Mix training rarely improves the coarse-grained OOD detection rates.

MixOE v.s. Mix + OE. Finally, we investigate whether naively combining the Mix and OE objective together can achieve similar effect to MixOE. A detailed analysis is given in Appendix D. The takeaway here is that Mix + OE will lead to manifold intrusion [8], where the training will assign distinct targets to inputs that are close to each other

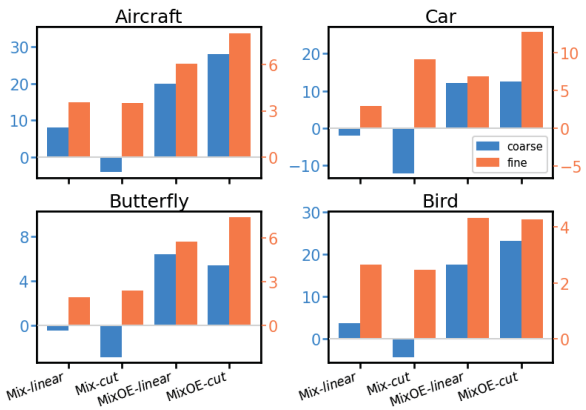

Figure 5. Comparison between MixOE and vanilla Mix training (without using outlier data) in terms of average difference in TNR95 relative to MSP. MixOE outperforms Mix against both coarse-/fine-grained OOD data.

in the DNN's feature space, causing significant learning difficulty for the model. Indeed, we find that when combining Mix and OE together, the model's accuracy can decrease by up to 10%, and the TNR95 can be worse than MSP by 10% and 20% against coarse- and fine-grained OOD samples, respectively. The results clearly demonstrate that *MixOE's formulation is unique, effective, and cannot be replaced by a simple/naive combination of two existing objectives.*

6. Conclusion

In this work, we propose Mixture Outlier Exposure, a DNN training algorithm for OOD detection in fine-grained environments. MixOE explicitly expand the coverage over the broad OOD region by mixing ID data and training outlier samples. The mixed samples are used to regularize the model's behaviour such that the prediction confidence smoothly decays when the inputs transition from ID to OOD. Experimental results in the four newly constructed large-scale fine-grained environments demonstrate that MixOE is able to improve detection rates against both coarse- and fine-grained OOD samples, while other methods hardly help with fine-grained detection. We hope that this work will facilitate and inspire future research on OOD detection in the challenging fine-grained settings.

Acknowledgement

This work is supported by FA8750-21-1-1015 and NSF-2140247.

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.

[2] Jiefeng Chen, Yixuan Li, Xi Wu, Yingyu Liang, and Somesh Jha. Atom: Robustifying out-of-distribution detection using outlier mining. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, pages 430–445. Springer, 2021.

[3] Tianshui Chen, Wenxi Wu, Yuefang Gao, Le Dong, Xiaonan Luo, and Liang Lin. Fine-grained representation learning and recognition by exploiting hierarchical semantic embedding. In *Proceedings of the 26th ACM international conference on Multimedia*, pages 2023–2031, 2018.

[4] Sanghyuk Chun, Seong Joon Oh, Sangdoo Yun, Dongyoon Han, Junsuk Choe, and Youngjoon Yoo. An empirical evaluation on robustness and uncertainty of regularization methods. *arXiv preprint arXiv:2003.03879*, 2020.

[5] 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*, pages 248–255. Ieee, 2009.

[6] Xuefeng Du, Zhaoning Wang, Mu Cai, and Sharon Li. Towards unknown-aware learning with virtual outlier synthesis. In *International Conference on Learning Representations*, 2022.

[7] Mengran Fan, Tapabrata Chakraborti, I Eric, Chao Chang, Yan Xu, and Jens Rittscher. Microscopic fine-grained instance classification through deep attention. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 490–499. Springer, 2020.

[8] Hongyu Guo, Yongyi Mao, and Richong Zhang. Mixup as locally linear out-of-manifold regularization. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 3714–3722, 2019.

[9] 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.

[10] Dan Hendrycks, Steven Basart, Mantas Mazeika, Mohammadreza Mostajabi, Jacob Steinhardt, and Dawn Song. Scaling out-of-distribution detection for real-world settings. *arXiv preprint arXiv:1911.11132*, 2019.

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

[12] Dan Hendrycks, Mantas Mazeika, and Thomas Dietterich. Deep anomaly detection with outlier exposure. In *International Conference on Learning Representations*, 2018.

[13] 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.

[14] 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*, pages 10951–10960, 2020.

[15] Rui Huang, Andrew Geng, and Yixuan Li. On the importance of gradients for detecting distributional shifts in the wild. *arXiv preprint arXiv:2110.00218*, 2021.

[16] Nathan Inkawhich, Eric Davis, Matthew Inkawhich, Uttam Majumder, and Yiran Chen. Training sar-atr models for reliable operation in open-world environments. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 14:3954–3966, 2021.

[17] Ramneet Kaur, Susmit Jha, Anirban Roy, Oleg Sokolsky, and Insup Lee. Are all outliers alike? on understanding the diversity of outliers for detecting oods. *arXiv preprint arXiv:2103.12628*, 2021.

[18] Umar Khalid. The method is not scalable., 2022.

[19] Aditya Khosla, Nityananda Jayadevaprakash, Bangpeng Yao, and Fei-Fei Li. Novel dataset for fine-grained image categorization: Stanford dogs. In *Proc. CVPR Workshop on Fine-Grained Visual Categorization (FGVC)*, volume 2. Citeseer, 2011.

[20] Jonathan Krause, Michael Stark, Jia Deng, and Li Fei-Fei. 3d object representations for fine-grained categorization. In *4th International IEEE Workshop on 3D Representation and Recognition (3dRR-13)*, Sydney, Australia, 2013.

[21] Kimin Lee, Honglak Lee, Kibok Lee, and Jinwoo Shin. Training confidence-calibrated classifiers for detecting out-of-distribution samples. In *International Conference on Learning Representations*, 2018.

[22] Kimin Lee, Kibok Lee, Honglak Lee, and Jinwoo Shin. A simple unified framework for detecting out-of-distribution samples and adversarial attacks. In *NeurIPS*, 2018.

[23] Qing Li, Weidong Cai, Xiaogang Wang, Yun Zhou, David Dagan Feng, and Mei Chen. Medical image classification with convolutional neural network. In *2014 13th international conference on control automation robotics & vision (ICARCV)*, pages 844–848. IEEE, 2014.

[24] Wen Li, Limin Wang, Wei Li, Eirikur Agustsson, and Luc Van Gool. Webvision database: Visual learning and understanding from web data. *arXiv preprint arXiv:1708.02862*, 2017.

[25] Shiyu Liang, Yixuan Li, and R Srikant. Enhancing the reliability of out-of-distribution image detection in neural networks. In *International Conference on Learning Representations*, 2018.

[26] Weitang Liu, Xiaoyun Wang, John Owens, and Yixuan Li. Energy-based out-of-distribution detection. *Advances in Neural Information Processing Systems*, 33, 2020.

[27] Ilya Loshchilov and Frank Hutter. Sgdr: Stochastic gradient descent with warm restarts. *arXiv preprint arXiv:1608.03983*, 2016.

[28] S. Maji, J. Kannala, E. Rahtu, M. Blaschko, and A. Vedaldi. Fine-grained visual classification of aircraft. Technical report, 2013.

[29] George A Miller. Wordnet: a lexical database for english. *Communications of the ACM*, 38(11):39–41, 1995.

[30] Tianyu Pang, Kun Xu, Yinpeng Dong, Chao Du, Ning Chen, and Jun Zhu. Rethinking softmax cross-entropy loss for ad-

versarial robustness. In *International Conference on Learning Representations*, 2019.

[31] 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*, pages 11544–11552, 2019.

[32] Alec Radford, Luke Metz, and Soumith Chintala. Unsupervised representation learning with deep convolutional generative adversarial networks. *arXiv preprint arXiv:1511.06434*, 2015.

[33] Deepak Ravikumar, Sangamesh Kodge, Isha Garg, and Kaushik Roy. Exploring vicinal risk minimization for lightweight out-of-distribution detection. *arXiv preprint arXiv:2012.08398*, 2020.

[34] Ryne Roady, Tyler L Hayes, Ronald Kemker, Ayesha Gonzales, and Christopher Kanan. Are open set classification methods effective on large-scale datasets? *Plos one*, 15(9):e0238302, 2020.

[35] Shibani Santurkar, Dimitris Tsipras, and Aleksander Madry. {BREEDS}: Benchmarks for subpopulation shift. In *International Conference on Learning Representations*, 2021.

[36] Chandramouli Shama Sastry and Sageev Oore. Detecting out-of-distribution examples with gram matrices. In *International Conference on Machine Learning*, pages 8491–8501. PMLR, 2020.

[37] 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.

[38] Engkarat Techapanurak, Anh-Chuong Dang, and Takayuki Okatani. Bridging in-and out-of-distribution samples for their better discriminability. *arXiv preprint arXiv:2101.02500*, 2021.

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

[40] Sunil Thulasidasan, Gopinath Chennupati, Jeff A Bilmes, Tanmoy Bhattacharya, and Sarah Michalak. On mixup training: Improved calibration and predictive uncertainty for deep neural networks. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019.

[41] Grant Van Horn, Steve Branson, Ryan Farrell, Scott Haber, Jessie Barry, Panos Ipeirotis, Pietro Perona, and Serge Belongie. Building a bird recognition app and large scale dataset with citizen scientists: The fine print in fine-grained dataset collection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 595–604, 2015.

[42] C. Wah, S. Branson, P. Welinder, P. Perona, and S. Belongie. The Caltech-UCSD Birds-200-2011 Dataset. Technical Report CNS-TR-2011-001, California Institute of Technology, 2011.

[43] Jim Winkens, Rudy Bunel, Abhijit Guha Roy, Robert Stanforth, Vivek Natarajan, Joseph R Ledsam, Patricia MacWilliams, Pushmeet Kohli, Alan Karthikesalingam, Simon Kohl, et al. Contrastive training for improved out-of-distribution detection. *arXiv preprint arXiv:2007.05566*, 2020.

[44] Gui-Song Xia, Xiang Bai, Jian Ding, Zhen Zhu, Serge Belongie, Jiebo Luo, Mihai Datcu, Marcello Pelillo, and Liangpei Zhang. Dota: A large-scale dataset for object detection in aerial images. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3974–3983, 2018.

[45] Gui-Song Xia, Jingwen Hu, Fan Hu, Baoguang Shi, Xiang Bai, Yanfei Zhong, Liangpei Zhang, and Xiaoqiang Lu. Aid: A benchmark data set for performance evaluation of aerial scene classification. *IEEE Transactions on Geoscience and Remote Sensing*, 55(7):3965–3981, 2017.

[46] 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.

[47] 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 *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 6023–6032, 2019.

[48] Hongyi Zhang, Moustapha Cisse, Yann N. Dauphin, and David Lopez-Paz. mixup: Beyond empirical risk minimization. In *International Conference on Learning Representations*, 2018.

[49] 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 international conference on computer vision*, pages 5209–5217, 2017.

[50] Xiao Xiang Zhu, Devis Tuia, Lichao Mou, Gui-Song Xia, Liangpei Zhang, Feng Xu, and Friedrich Fraundorfer. Deep learning in remote sensing: A comprehensive review and list of resources. *IEEE Geoscience and Remote Sensing Magazine*, 5(4):8–36, 2017.