From Coarse to Fine-Grained Open-Set Recognition

Nico Lang[1] Vésteinn Snæbjarnarson[1] Elijah Cole[2]
Oisin Mac Aodha[3] Christian Igel[1] Serge Belongie[1]
[1]University of Copenhagen [2]Altos Labs [3]University of Edinburgh

Abstract

Open-set recognition (OSR) methods aim to identify whether or not a test example belongs to a category observed during training. Depending on how visually similar a test example is to the training categories, the OSR task can be easy or extremely challenging. However, the vast majority of previous work has studied OSR in the presence of large, coarse-grained semantic shifts. In contrast, many real-world problems are inherently fine-grained, which means that test examples may be highly visually similar to the training categories. Motivated by this observation, we investigate three aspects of OSR: label granularity, similarity between the open- and closed-sets, and the role of hierarchical supervision during training. To study these dimensions, we curate new open-set splits of a large fine-grained visual categorization dataset. Our analysis results in several interesting findings, including: (i) the best OSR method to use is heavily dependent on the degree of semantic shift present, and (ii) hierarchical representation learning can improve coarse-grained OSR, but has little effect on fine-grained OSR performance. To further enhance fine-grained OSR performance, we propose a hierarchy-adversarial learning method to discourage hierarchical structure in the representation space, which results in a perhaps counter-intuitive behaviour, and a relative improvement in fine-grained OSR of up to 2% in AUROC and 7% in AUPR over standard training. Code and data are available: langnico.github.io/fine-grained-osr.

1. Introduction

The goal of *open-set recognition* (OSR) is to distinguish *familiar* or *closed-set* categories that were seen during training from *novel* or *open-set* categories that were not seen [45]. In other words, OSR focuses on detecting semantic shifts between the training and test categories.[1] However, not all semantic shifts are equal. Different open-set categories will have different degrees of visual similarity to the familiar closed-set categories. Recently, Vaze et al. [53] showed that OSR is harder for categories exhibiting *smaller* semantic shifts relative to the training categories. Intuitively, novel categories that are highly similar to the training categories tend to be confused with familiar ones with high confidence. We refer to this phenomenon as the "familiarity trap". Note that this is the opposite of what we expect to see in zero-shot classification [42] and transfer-learning [12], where smaller semantic shifts make the problem easier, not harder (see illustration in Figure 1). The vast majority of the OSR literature has focused on investigating *coarse* open-set splits [53] while the impact of finer-grained splits is underexplored.

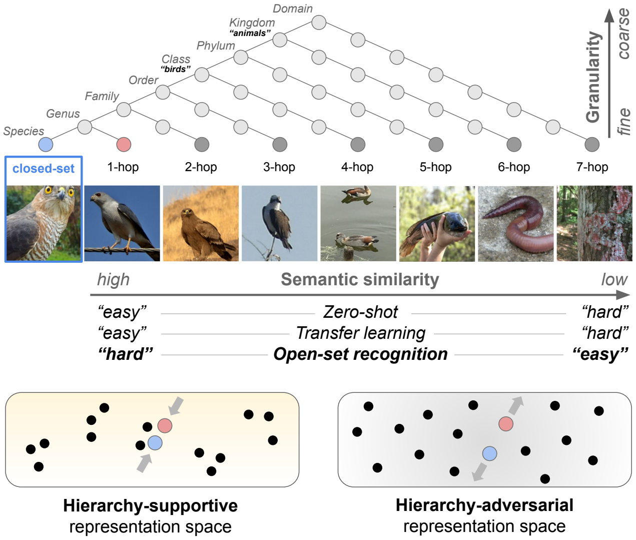


Figure 1. **Granularity and semantic similarity are understudied confounders in open-set recognition (OSR).** Given a taxonomic label hierarchy, we curate open-set splits with increasing semantic shifts (i.e., hops) to show that careful attention to these factors is instrumental in answering questions such as: What is the best OSR scoring method? Do hierarchical representations improve OSR?

Unlike traditional (coarse-grained) OSR, fine-grained OSR reflects realistic category discovery challenges in fine-

[1] In this work, we use the taxonomic distance between categories in a label hierarchy to quantify semantic dissimilarity. We assume that this distance correlates with visual dissimilarity between the categories.

grained domains like medicine [43], art [11], and the natural world [51]. For example, in species monitoring [3, 27, 39, 51], we might expect to see new species that are highly visually similar to known species. While existing unlabeled data might contain such examples, their similarity makes it difficult to detect them. In citizen science applications, automated vision systems can be used to distinguish poisonous mushrooms from similar-looking edible ones [41]. Ensuring that a novel, potentially dangerous mushroom species is not confused with a previously observed safe one is crucial. Fine-grained OSR problems may also be encountered when digitizing museum collections, screening potentially counterfeit goods, or identifying fashion trends.

Natural world image collections from citizen science platforms like iNaturalist [1] are particularly well-suited to studying fine-grained OSR thanks to their size, diversity, and high-quality taxonomic structure (which provides a proxy for semantic similarity). In this work, we use the iNat2021 [52] dataset to investigate the challenges in fine-grained OSR in terms of semantic similarity, supervision granularity, and hierarchical representations. We explore the role of semantic shift by studying progressively more difficult open-set splits (Figure 1). In contrast to previous work that uses fine-grained datasets with a limited number of categories [53], we work with a large-scale benchmark containing 10,000 categories and 2.7 million images. This allows us to study diverse real-world OSR phenomena. By training models from scratch on large splits of iNat2021, we avoid the risk of leaking knowledge from pretraining datasets like ImageNet [44] that might contain semantic overlap with the open-set [7].

While the use of hierarchical labels has been studied in closed-set recognition to reduce the severity of mistakes [6, 15, 17] in fashion [31], crop mapping [50], food [58] and species recognition [14], the role of hierarchical representations for fine-grained OSR is underexplored. We find that encouraging hierarchical structure in the representation space can be beneficial for closed-set classification and coarse-grained open-set recognition, but surprisingly it has a limited effect on fine-grained OSR. Motivated by this finding, we explore whether implicit hierarchical structure should be reduced for fine-grained OSR. We propose a *hierarchy-adversarial learning* approach that can improve fine-grained OSR performance by discouraging hierarchical structure. By adapting the gradient reversal layer developed for unsupervised domain adaptation [20, 21], our approach learns representations that reward linear separability of fine-grained categories while discouraging linear separability of coarser-grained granularities. As we discourage hierarchical structure, representations of instances from categories that share similar features are pushed away from each other. To put it differently, we hypothesize that the hierarchy-adversarial approach discourages features shared across classes in favour of class-specific features. Thus, fine-grained novel categories are less likely to get caught by the "familiarity trap", i.e., less likely to be misclassified as a semantically similar category with high confidence.

We make the following contributions:

1. We demonstrate that the choice of the best scoring rule for OSR depends on the *semantic similarity* between the closed- and open-set and that familiarity scores perform best in fine-grained OSR.
2. We investigate the impact of *supervision granularity* and show that fine-grained supervision improves OSR performance even for large semantic shifts.
3. We explore the role of hierarchical representations and introduce *hierarchy-adversarial learning* to improve fine-grained OSR by discouraging hierarchical structure.
4. We introduce iNat2021-OSR, a benchmark with curated open-set splits for the iNat2021 dataset [52] for two taxa: birds and insects. This enables the study of OSR along seven discrete "hops" that encode the semantic distance from coarse-grained (7-hop) to fine-grained (1-hop).

2. Related work

2.1. Open-set recognition

Open-set recognition (OSR) [45] is concerned with the problem of identifying novel categories, i.e., semantic shifts. A large number of OSR-specific deep learning methods have been proposed including OpenMax [4], OSRCI [37], ARPL [8], and OpenHybrid [60]. However, recently Vaze et al. [53] demonstrated that simple baselines based on classifier confidence (i.e. logits) perform on par and that the improvement in OSR performance is not necessarily solely attributed to dedicated OSR approaches, but can instead be explained by improvements in the closed-set recognition accuracy, as the result of using more sophisticated deep neural network architectures and training strategies. In parallel, Dietterich and Guyer [18] proposed the 'familiarity hypothesis' to explain the observation that deep networks can be interpreted as detecting the *absence* of familiarity, instead of the *presence* of novelty.

In this work, we are concerned with *fine-grained* OSR, a setting which has been partially explored in existing work [2, 13, 22, 48, 53]. Our goal is to investigate the impact in performance resulting from changes in the semantic similarity between the closed and open-sets. Many existing fine-grained approaches construct random data splits, which will thus contain a mixture of hard and easy categories in the open-set [22], or they use splits with very large semantic differences (e.g., using dogs as familiar and cars as novel categories) Dai et al. [13]. Using manually annotated attributes, Vaze et al. [53] constructed easy, medium, and hard open-set splits of three popular fine-grained datasets (i.e., CUB-Birds [55], Stanford Cars [30], and FGVC-Aircraft [36]). We build on this work, by constructing dataset splits leveraging taxonomies to allow us to explore the impact of a larger variety of split types.

There are two major limitations in prior OSR work that can yield an overly optimistic impression of model performance. First, the small size of existing OSR benchmark datasets necessitates the use of pretrained models. This potentially results in a semantic overlap between the pretraining categories (e.g., from ImageNet [44]) and the open-set categories [25]. Hence, the open-set categories may already be well separated from the closed-set ones. Second, the experimental evaluation on random data splits can lead to *coarse* open-set splits, which are not representative of real-world open-set applications that can be finer-grained [53]. This work aims to overcome both limitations by conducting experiments on large-scale image datasets across a spectrum of coarse to fine open-set splits.

Related to OSR is the more general task of out-of-distribution (OOD) detection, also referred to as anomaly or novelty detection [33, 47], which encompasses a set of methods that explore problems such as detecting covariate or distributional shifts [23, 29, 40, 62], semantic shifts [25], and combinations of shifts [26, 40, 49]. OSR is concerned with semantic shifts, and can thus be interpreted as a particular case of OOD detection [59]. Different methods have been explored for OOD detection, including score-based [24–26, 34], ensemble/disagreement-based [32, 32, 40], and distance metric-based [28]. In Sec. 5, we evaluate these three families of methods in the context of OSR.

2.2. Granularity, similarity, and hierarchy

Our work is concerned with exploring OSR through the lens of granularity, similarity, and hierarchy, we briefly summarize relevant related works on these topics below.

Label granularity encodes how detailed a category label is, and can range from coarse-grained (e.g., "animal") to fine-grained (e.g., "blue jay"). Granularity has been widely studied in closed-set recognition [57]. Multiple granularity supervision has been used to implicitly focus on different parts of an image to improve fine-grained categorization [56]. The granularity of supervision can impact performance in different ways.Van Horn et al. [51] showed that fine-grained training labels can hurt performance on coarse-grained object detection. Cole et al. [9] demonstrated that there is a "sweet" spot of granularity supervision for weakly supervised object localization (i.e., not too fine and not too coarse). Furthermore, it has been demonstrated that the accuracy gap between self-supervised and supervised methods grows when evaluated using finer-grained labels [10].

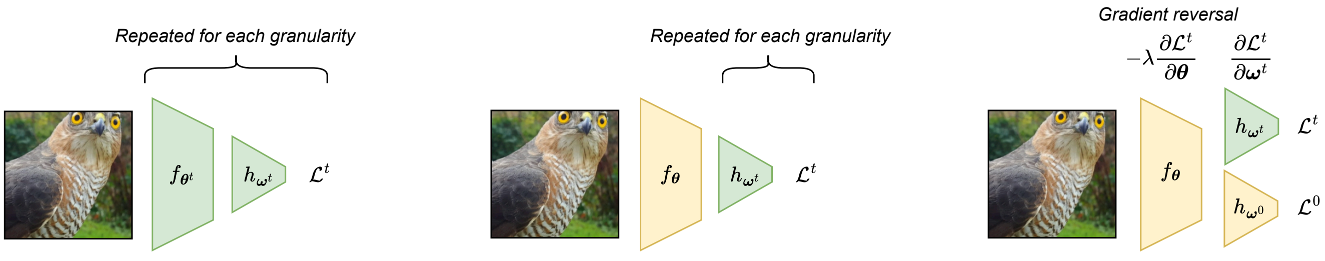
Semantic similarity describes the distance between two concepts (e.g., categories) in terms of their meaning. Visual similarity (or perceptual similarity [61]) can serve as a proxy for semantic similarity. Selecting visually similar pretraining data has been shown to be effective for improving transfer learning performance [12]. In fine-grained domains, samples of the same category can be visually very different (i.e., exhibit high intra-class variance) and samples of different categories can be very visually similar (i.e., exhibit low inter-class variance) [57]. Depending on the domain, nuisance variables such as object pose and scene illumination [46, 62], or sex and age [51] can amplify the discrepancy between visual and semantic similarity. Additional metadata such as attributes or taxonomy has been studied as an alternative proxy for semantic similarity, for instance in zero-shot learning in ImageNet categories [19, 38] or for iNaturalist species [42]. Vaze et al. [53] used labeled attributes to demonstrate that open-set recognition performance depends on the semantic similarity between the training and test categories. Interestingly, the difficulty of a task is not consistently coupled with granularity or semantic similarity. While OSR becomes harder for small semantic shifts, in contrast, zero-shot and transfer learning becomes easier.

Label hierarchy has been incorporated in closed-set recognition to reduce the severity of mistakes [6, 15, 17] in fashion [31], crop mapping [50], food [58] and species recognition [14]. Similarly, in hierarchical OSR [5, 16, 33, 47, 54] the aim is to find the closest ancestor, the 'super-category' of a novel category. This task turns the open-set problem into a closed-set one by coarsening the granularity of the predicted category. Our goal is not to advance hierarchical OSR, but to instead study the role of granularity, semantic similarity, and hierarchical structure in OSR, with the aim of understanding how to advance performance in the presence of very small semantic shifts, i.e., *fine-grained OSR*.

3. Methodology

3.1. Scores for open-set recognition

For the task of binary OSR, i.e., familiar *vs*. novel, we model a score $S(\boldsymbol{x})$ to indicate if a test sample $\boldsymbol{x}$ belongs to the set of familiar training categories $\mathcal{F} = \{1, \ldots, K\}$. The score $S(\boldsymbol{x})$ should increase as $p(y \in \mathcal{F}|\boldsymbol{x})$ increases, i.e., there should be a high rank correlation between the two. We summarize three score methods from the OOD and OSR literature that cover three families of methods [32, 49, 53]. For a given input image, we extract representations using the encoder $f_{\boldsymbol{\theta}}$, e.g., a trained deep neural network, parameterized by $\boldsymbol{\theta}$. A linear decoding function $h_{\boldsymbol{\omega}}$ with parameters $\boldsymbol{\omega}$ maps the representations to a K-dimensional vector, the logits, and the softmax function σ maps these logits to probabilities. That is, the predicted probability of an input image $\boldsymbol{x}$ belonging to class y is given by $p(y|\boldsymbol{x};\boldsymbol{\theta},\boldsymbol{\omega}) = \sigma_y\left(h_{\boldsymbol{\omega}}\left(f_{\boldsymbol{\theta}}\left(\boldsymbol{x}\right)\right)\right)$. We write $p(y|\boldsymbol{x})$ for $p(y|\boldsymbol{x};\boldsymbol{\theta},\boldsymbol{\omega})$ if the model parameters are clear from the context.


Repeated for each granularity
f_{θ^t}
h_{ω^t}
$\mathcal{L}^t$
Repeated for each granularity
$f_{\boldsymbol{\theta}}$
h_{ω^t}
$\mathcal{L}^t$
Gradient reversal
$-\lambda \frac{\partial \mathcal{L}^t}{\partial \boldsymbol{\theta}}$
$\frac{\partial \mathcal{L}^t}{\partial \boldsymbol{\omega}^t}$
$f_{\boldsymbol{\theta}}$
h_{ω^t}
$\mathcal{L}^t$
h_{ω^0}
$\mathcal{L}^0$

(a) Supervision granularity

(b) Hierarchy-supportive

(c) Hierarchy-adversarial

Figure 2. **Training strategies.** We train and compare learning strategies with (a) separate models for each level t in a hierarchy (*supervision granularity*), (b) single models that predict all levels of granularities (*hierarchy-supportive*), and (c) single models that discourage prediction capabilities of all but the finest granularity (*hierarchy-adversarial*). To discourage hierarchical structure in the representations, we use a gradient reversal layer from unsupervised domain adaptation [20] that reverses gradients for the classification heads of coarser granularities, i.e., $t > 0$.

Familiarity scores. These methods use either the logits of a classifier or the softmax confidence scores. It has been empirically shown that the unnormalized maximum logit score (MLS) can outperform the maximum softmax probability (MSP) as an open-set scoring function [53]. The MLS score is defined as:

$$S_{\text{MLS}}(\boldsymbol{x}) \triangleq \max_{y \in \mathcal{F}} [h_{\boldsymbol{\omega}} (f_{\boldsymbol{\theta}} (\boldsymbol{x}))]_y, \tag{1}$$

where $[\cdot]_y$ denotes the y-th component of a logit vector.

Disagreement scores. We consider an ensemble of M classifiers with parameter $\boldsymbol{\theta}_m$ and $\boldsymbol{\omega}_m$ for $m = 1, \ldots, M$, and write $p_m(y|\boldsymbol{x}) = p(y|\boldsymbol{x}; \boldsymbol{\theta}_m, \boldsymbol{\omega}_m)$. We can compute not only the average confidence $p_E(y|\boldsymbol{x}) = \frac{1}{M} \sum_m p_m(y|\boldsymbol{x})$ and average logits $l_E(y|\boldsymbol{x}) = \frac{1}{M} \sum_m h_{\boldsymbol{\omega}_m} (f_{\boldsymbol{\theta}_m} (\boldsymbol{x}))$, but also an alternative score that quantifies the disagreement (i.e., variance) between the individual member outputs. The negative KL-disagreement [32], S_{KLD} is used as an indication that a high score denotes familiar categories:

$$S_{\text{KLD}}(\boldsymbol{x}) \triangleq - \sum_{m=1}^{M} \text{KL} \left(p_E(\cdot|\boldsymbol{x}) \| p_m(\cdot|\boldsymbol{x}) \right), \tag{2}$$

where KL represents the Kullback-Leibler divergence between the predicted class distribution of the ensemble $p_E(y|\boldsymbol{x})$, i.e., the average confidence over the M members, and the predicted class distribution of each member $p_m(y|\boldsymbol{x})$.

Distance metric scores. Finally, distance in representation space can be combined with k-nearest neighbours (KNN) to formulate an OSR score. We adopt the nearest-neighbour approach from [49], originally proposed for general OOD detection. The method takes the L2-normalized representations of the encoder $\tilde{f}_\theta(\boldsymbol{x}) = f_{\boldsymbol{\theta}}(\boldsymbol{x}) / \| f_{\boldsymbol{\theta}}(\boldsymbol{x}) \|_2$ and computes the shortest distance to the images in the training data set $\{(\boldsymbol{x}_1, y_1), \ldots, (\boldsymbol{x}_N, y_N)\}$ in the normalized representation space:

$$S_{\text{NN}}(\boldsymbol{x}) \triangleq \min_{i=1,\ldots,N} - \| \tilde{f}_\theta(\boldsymbol{x}) - \tilde{f}_\theta(\boldsymbol{x}_i) \|_2. \tag{3}$$

3.2. Hierarchy-supportive learning

One of our objectives is to study the role of label training granularity and hierarchy in OSR. Among many existing approaches [6, 14, 15, 17, 31, 50], we chose a simple hierarchical multi-task approach [58]. This provides us with a framework for exploring representations where the hierarchical structure is either encouraged or discouraged. We follow the notation in [58] and define a hierarchy of a given label space as $\mathcal{T} = \{\mathcal{Y}^t\}_{t=0}^{T}$, where $\mathcal{Y}^t$ defines the set of categories at the t-th level in the hierarchy and the leaf nodes at $t = 0$ correspond to the fine-grained categories. From this, we can formulate a hierarchy-supportive multi-task loss that optimizes the cross-entropy at every level t in the hierarchy, where y_i^t denotes the true label for sample i at the t-th hierarchy level and $p^t(y_i^t|\boldsymbol{x}_i) = p(y_i^t|\boldsymbol{x}_i; \boldsymbol{\theta}, \boldsymbol{\omega}^t)$:

$$\mathcal{L}_{\text{HS}} = \sum_{t=0}^{T} \sum_{i=1}^{N} - \log p^t(y_i^t|\boldsymbol{x}_i). \tag{4}$$

To optimize this training loss, we append additional classification heads $h_{\boldsymbol{\omega}^t}$ with parameters $\boldsymbol{\omega}^t$ to the encoder, one for each level t in the hierarchy as illustrated in Figure 2b.

3.3. Hierarchy-adversarial learning

Similarly, we can formulate a multi-task loss for fine-grained classification that *discourages* hierarchical structure in the representation space. We are interested in learning an encoder $f_{\boldsymbol{\theta}}$ that yields a representation suitable for classifying fine-grained categories – but is agnostic to coarser hierarchical labels at levels $t > 0$ in the hierarchy. We hypothesize that this encourages salient features specific for fine-grained categories rather than relying on coarse-grained features for distinguishing groups of fine-grained categories. This can be achieved by minimizing the cross-entropy for the finest granularity $\mathcal{L}^0$, but maximizing the cross-entropy for the coarser granularities $\mathcal{L}^t$ with $t > 0$,

$$\mathcal{L}_{\text{HA}} = \underbrace{\sum_{i=1}^{N} - \log p^0(y_i^0|\boldsymbol{x}_i)}_{\mathcal{L}^0} - \lambda \sum_{t=1}^{T} \underbrace{\sum_{i=1}^{N} - \log p^t(y_i^t|\boldsymbol{x}_i)}_{\mathcal{L}^t} . \tag{5}$$

To ensure that the $\mathcal{L}^t$ for $t > 0$ are not maximized by the decoding functions, but by the representation, we solve a min-max optimization problem:

$$\min_{\boldsymbol{\theta}, \boldsymbol{\omega}^0} \max_{\boldsymbol{\omega}^1, \ldots, \boldsymbol{\omega}^T} \mathcal{L}_{\text{HA}}(\boldsymbol{\theta}, \boldsymbol{\omega}^0, \boldsymbol{\omega}^1, \ldots, \boldsymbol{\omega}^T). \tag{6}$$

That is, we obtain the parameters of the encoder $\boldsymbol{\theta}$ and the fine-grained decoder $\boldsymbol{\omega}^0$ that *minimize* $\mathcal{L}^0$ (i.e., the representation is discriminative for fine-grained labels). At the same time, for $t > 0$, we obtain the decoders $\boldsymbol{\omega}^t$ that *maximize* $\mathcal{L}^t$, i.e., the representation is regularized to not capture coarser-grained hierarchical labels. To train a deep neural network with this min-max objective using stochastic gradient descent (SGD), we adapt a *gradient reversal layer*-based approach inspired by the domain adaptation literature [20, 21]. By swapping the sign of the gradient corresponding to the $\mathcal{L}^t$ terms before back-propagating into the encoder (see Figure 2c), all learnable parameters can be updated in an otherwise standard forward- and backward-pass.

4. Experimental setup

4.1. The iNat2021-OSR dataset

While existing OSR benchmarks for coarse-grained labels contain many samples, the fine-grained ones are small in size [53]. This limitation of the fine-grained datasets makes it impossible to train models from scratch on them, which we require to have full control over a realistically plausible setting for OSR. We are also interested in having a granular label hierarchy that can serve as a proxy for semantic similarity. Motivated by this, we look to the iNat2021 dataset [52] which contains over 2.5 million images of 10,000 different plant and animal species (i.e., categories) as a benchmark for fine-grained OSR.

Our experimental setup is inspired by the zero-shot learning work of Rodríguez et al. [42], who use the taxonomy of iNat2021 to evaluate zero-shot performance at different semantic distances to the training data, measured in the number of edges to the lowest common ancestor in the hierarchical graph, we simply refer to this distance as *hops* (see Figure 1). We develop two curated open-set data splits of iNat2021 (see Table 1) and group open-set categories by the smallest hop distance to any sample in the training data, ranging from 1–7 hops. The two splits provide a closed-set for the super-categories “birds” (Aves in biological classification) and “insects” (Insecta) at the hierarchy level $t = 5$ (i.e., “class”). The closed-set consists of a subset of species from the respective super-category, i.e., the familiar categories. The remaining categories are used as the open-set test sets with semantic distances ranging from 1-hop (fine-grained OSR) to 7-hop (coarse-grained OSR). While hops 1–4 correspond to relationships within the super-categories (i.e., novel bird or insect species), hops > 4 correspond to open-set categories outside these super-categories. Full details about the iNat2021-OSR dataset are given in the supplementary material.

	Train	Test	Open-set test						
	familiar	familiar	1-hop	2-hop	3-hop	4-hop	5-hop	6-hop	7-hop
Aves									
Categories	745	745	297	180	170	94	930	2972	4607
Samples	210323	7450	2970	1800	1700	940	9300	29720	46070
Insecta									
Categories	1501	1501	505	333	99	88	226	2636	4587
Samples	398952	15010	5050	3330	990	880	2260	26360	45870

Table 1. **iNat2021-OSR data split statistics.** Based on the iNat2021 benchmark dataset [52], we curate two new large-scale OSR benchmark datasets for the super-categories “birds” (Aves) and “insects” (Insecta). Both versions contain seven open-set splits of increasing semantic distance (fine to coarse).

4.2. Implementation details

Training. We use a ResNet-50 backbone and train all models from scratch for 100 epochs using SGD with a base learning rate set to 0.1, which is multiplied by factor 0.1 every 30 epochs, and weight decay of $1\mathrm{e}{-4}$. We train five models for the ensemble disagreement score (Sec. 3.1) starting from different random initializations following prior work [40]. To train the hierarchy-adversarial approach (Sec. 3.3), we gradually increase the weight of the adversarial gradient λ starting from zero to α by the end of the training process similar to [20].

Evaluation. Open-set performance is reported using an equal number of samples in the closed-set and open-set by sub-sampling the larger set. For all evaluation metrics, a higher score means better performance. More details can be found in the supplementary material. Closed-set accuracy of coarser granularities is evaluated using hard pooling, based on a single predicted category and the given hierarchy. We did not observe a difference using soft pooling, i.e., aggregating the probabilities of multiple categories.

5. Results

Our experiments investigate the role of *semantic similarity*, *supervision granularity*, and *hierarchical structure* in OSR.

5.1. What is the role of semantic similarity in OSR?

Fine-grained OSR is hardest. We observe a large performance gap between fine-grained (1-hop) and coarse-grained OSR (7-hop) (Figure 3). Hence, OSR performance correlates positively with semantic distance. One exception is the performance drop from 6-hop to 7-hop, where the

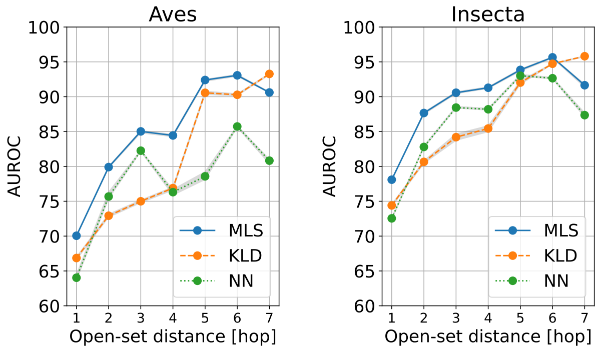


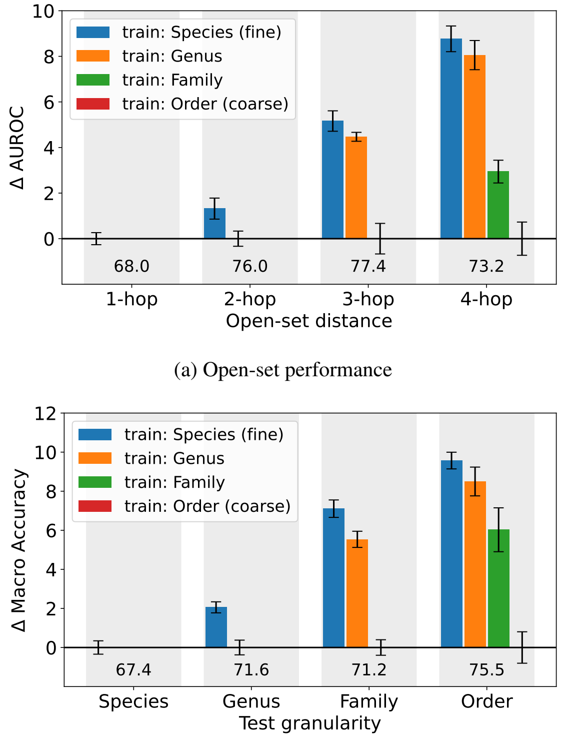
Figure 3. **Comparison of OSR score methods.** MLS: Maximum logit score [53], KLD: KL-disagreement [32], and NN: Nearest neighbour [49]. OSR results for the MLS are averaged over 5 resampled ensembles trained on the finest granularity on iNat2021-OSR. Each ensemble consists of 5 randomly sampled members drawn without replacement from a pool of 10 models. The shaded area indicates the min and max AUROC.

familiar categories from "birds" and "insects" are tested against "plants" and "fungi". A possible explanation for this drop could be that plants are often seen during training in the background for the closed-set categories. This may increase familiarity scores, making these samples harder to detect as novel categories. Similar issues have been found to impair the evaluation of OOD detection [7].

The choice of OSR score depends on semantic similarity. All scores follow the same general trend that OSR performance improves with larger semantic distance (Figure 3). However, none of the scoring rules consistently performs best across the spectrum of semantic similarity. For fine-grained OSR (1-hop) and up to 6-hop, the maximum logit score (MLS) performs best. However, for coarse-grained OSR, the KL-disagreement outperforms the logit score in our experiments. Moreover, as birds and insects contain category relationships up to 4-hops, open-sets with hops >4 are completely outside the classifier's domain expertise. This explains the discontinuity from 4- to 5-hop, where the performance gap between the KL-disagreement and the logit score is reduced. The nearest neighbour approach performs worst for both extremes. Interestingly, the L2-distance on unnormalized representations yields a lower performance on 1-hop, but higher performance on 7-hop OSR (see supplementary Figure A6). As the MLS performs best in fine-grained OSR, we will focus our analyses on this score.

5.2. What is the impact of supervision granularity?

Closed-set accuracy increases for coarser granularities. By pooling both the fine-grained prediction and the target label to coarser granularities, the closed-set accuracy improves (Figure 4b). This confirms the observation that the closed-set problem is easier for coarser granularities with fewer categories [10].



(a) Open-set performance

(b) Closed-set performance

Figure 4. **Supervision granularity.** Performance for the Aves split when training on granularities that are finer than used for testing for both (a) open-set and (b) closed-set recognition. The reference at zero is the performance of the model trained on the same granularity as the test granularity (corresponding to open-set distance). For context, the performance of this baseline model is provided above the test granularity label. The error bars indicate the standard deviation over 5 training runs. For the closed-set, we report the macro accuracy (averaging per-class) to account for class imbalance in coarser granularities.

Overly fine-grained supervision improves closed-set accuracy. Closed-set accuracy is highest when we pool the model trained on the finest granularity to the coarser test granularity. Training on the matching granularity deteriorates performance (Figure 4b). This contradicts results in both object detection, where overly fine supervision deteriorates performance for coarser super-categories [51], and weakly supervised localization where a "sweet spot" of medium-grained supervision performs best [9]. The iNat2021 dataset is approximately class-balanced at the finest granularity, but is imbalanced at coarser levels as the number of leaf nodes varies in the hierarchy. Therefore, we average the per-category accuracies, which shows a substantial performance gap between the matching granularity and the pooling of the finest supervision granularity.

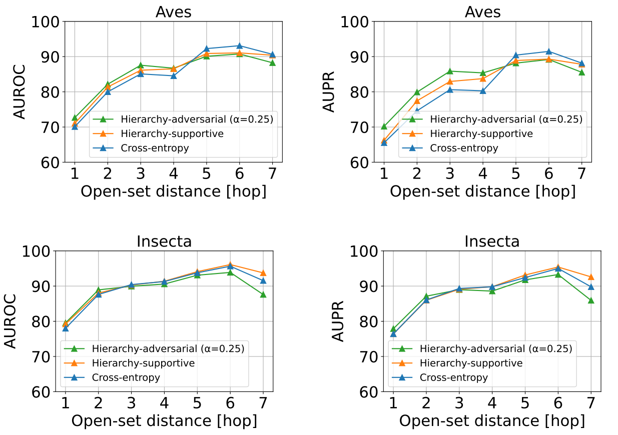


Figure 5. **Hierarchy-aware training strategies.** OSR results for Aves (top) and Insecta (bottom) using the maximum logit score (MLS) of ensembles with 5 models. We compare the training strategies: Cross-entropy on the fine-grained labels, hierarchy-supportive (Eq. 4) and hierarchy-adversarial (Eq. 5).

Overly fine-grained supervision improves OSR. Like the closed-set accuracy, OSR performance is higher for models trained on the finer granularities (Figure 4a). Even if the familiar and novel data is separated at coarser granularities, i.e., hop distances >1, training models on finer granularity is still beneficial. It is also best for finest-granularity supervision across the entire spectrum of shifts. This behaviour can be explained by the phenomena observed by Vaze et al. [53] whereby better closed-set accuracy improves OSR performance.

5.3. How does hierarchical structure affect OSR?

Hierarchical representations can improve coarse-grained OSR. First, we look at the closed-set accuracy when using hierarchy-supportive learning, which is unaffected at the finest granularity (Figure 5 and Table 2), likely because the coarse supervision does not help the learning of features to resolve confusions within very similar categories. However, closed-set accuracy improves for coarser granularities (see supplementary), which means that the severity of mistakes is reduced and is in line with previous work [6, 15, 17]. Furthermore, the hierarchical structure can also improve coarse-grained OSR performance (see Figure 5), which might be explained by the correlation between closed-set and open-set performance [53].

	Aves			Insecta		
Training strategy	ACC	AUROC	AUPR	ACC	AUROC	AUPR
Cross-entropy	67.4 (0.3)	68.0 (0.3)	64.8 (0.3)	**82.9** (0.3)	75.3 (0.3)	74.5 (0.3)
H-supportive	66.6 (0.4)	69.3 (0.5)	65.9 (0.5)	82.3 (0.1)	**76.6** (0.3)	74.9 (0.4)
H-adversarial (α=0.25)	**67.6** (0.3)	**69.9** (0.5)	**69.1** (0.5)	82.7 (0.3)	**76.6** (0.4)	**75.9** (0.5)

Table 2. **Fine-grained OSR (1-hop) and closed-set performance.** MLS for single models averaged over 5 models trained with cross-entropy and hierarchy-aware strategies. OSR performance (AUROC, AUPR) and the corresponding closed-set accuracy (ACC) are reported with standard deviations in parentheses.

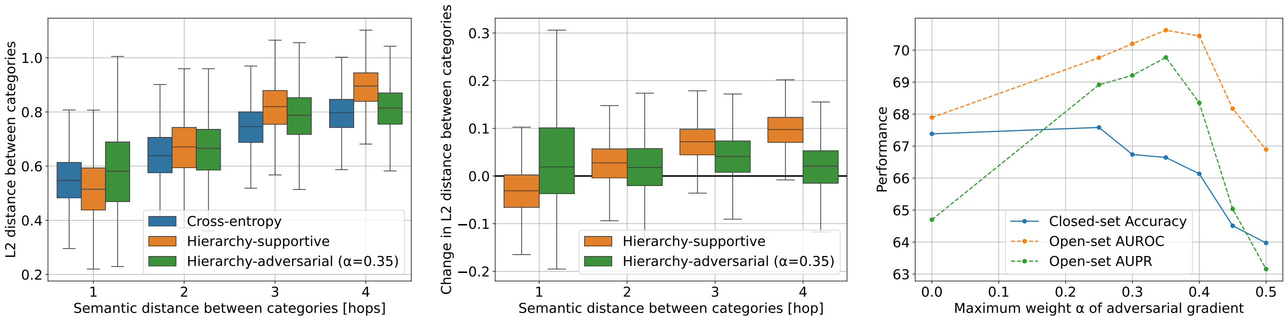
		Aves		Insecta	
Score	Training strategy	AUROC	AUPR	AUROC	AUPR
MLS	Cross-entropy	70.1	65.5	78.0	76.4
	Hierarchy-supportive	71.1	66.2	79.2	76.4
	Hierarchy-adversarial (α=0.25)	**72.7**	**70.2**	**79.5**	**77.9**
NN	Cross-entropy	64.1	61.7	72.6	68.0
	Hierarchy-supportive	65.6	60.8	74.0	68.2
	Hierarchy-adversarial (α=0.25)	**67.8**	**62.6**	**76.4**	**72.1**
KLD	Cross-entropy	**67.0**	**59.4**	**74.6**	**65.7**
	Hierarchy-supportive	66.0	58.2	73.7	65.3
	Hierarchy-adversarial (α=0.25)	66.1	59.1	72.0	63.8

Table 3. **Fine-grained OSR (1-hop) ensemble performance.** Results for ensembles with 5 models trained with cross-entropy and hierarchy-aware strategies. MLS: Maximum logit score [53], NN: Nearest neighbour [49], KLD: KL-disagreement [32].

Hierarchy-adversarial learning can improve fine-grained OSR. Based on these observations, we study an alternative direction and turn to the hierarchy-adversarial training strategy, which aims to reduce hierarchical structure in the learned representation. This approach improves fine-grained OSR for 1-hop open-sets across several metrics for both the MLS and NN score, but not for the KL-disagreement (see Figure 5 and Tables 2,3). However, it consistently impairs coarser OSR performance (Figure 5). Qualitative examples where the hierarchy-adversarial approach improves the "familiarity trap" w.r.t. the baseline are included in the supplementary material.

To empirically analyse this strategy's effect, we compute the L2-distance between category centroids from the 1-hop open-set categories and the training categories in the learned representation space (Figure 6 a,b). We observe that hierarchy-adversarial learning pushes fine-grained open-set categories away from their most related training categories (1-hop). This contrasts hierarchy-supportive learning, which pulls the fine-grained open-set closer to its most related training categories but pushes categories at larger distances even further away.

We study the sensitivity of the hyperparameter λ from Equation 5, which controls the weight of the adversarial gradient and gradually increases from zero to α. We demonstrate that this parameter controls both the closed-set accuracy and the open-set performance (Figure 6 c). First, by increasing α, we see an improvement in open-set performance, but at the cost of closed-set accuracy. This behaviour contradicts the observation from Vaze et al. [53] as it shows that there is an underlying role of hierarchical structure in OSR that cannot necessarily be explained by improved closed-set accuracy. However, open-set performance drops at a point, most likely due to the drop in closed-set accuracy.



(a) (b) (c)

Figure 6. **Hierarchy-adversarial learning analyses.** To analyze how hierarchy-aware training affects the relationship between classes in feature space, we plot (a) the L2-distance between novel and familiar category centroids in representation space and (b) the change of the L2-distance w.r.t. the cross-entropy baseline (derived from a). We expect that the *hierarchy-supportive* learning pulls fine-grained open-set categories (1-hop) closer to the familiar categories and amplifies the distance between coarse-grained categories. In contrast, the *hierarchy-adversarial* learning is expected to push fine-grained categories away from familiar categories, but roughly preserve the distance of coarse-grained categories. Our empirical analyses confirm these expectations on the Aves dataset. (c) Hyperparameter study for the weight of the hierarchy-adversarial gradient on the fine-grained open-set (1-hop). With increasing α, open-set recognition performance improves at the cost of closed-set accuracy. However, there is a tipping point when the weight α is too high, and open-set performance drops. This behaviour contradicts with the observation of Vaze et al. [53] as closed-set and open-set performance do not correlate.

5.4. Limitations

While our study presents a step towards understanding OSR under challenging fine-grained shifts, some limitations exist. First, we rely on a taxonomy as a proxy for concept granularity and semantic similarity to study our research questions. To fit our needs, we have curated a new large-scale open-set benchmark dataset using “birds” and “insects” from iNat2021 as the training domains. It remains to be tested if our observations hold for other domains, such as food, art, or medicine. However, Cole et al. [9] have demonstrated that the existing ImageNet hierarchy is not a good proxy for concept granularity. Thus, ImageNet is not a likely candidate for our OSR setting either. An additional challenge is the irregular depth of the ImageNet categories.

Second, controlling the degree of hierarchical structure in learned representations is non-trivial. Some hierarchical structure is essential to solving fine-grained problems and is learned implicitly, as shown in our experiments (see Figure 4). We study the “opposite” of hierarchical structure with a hierarchy-adversarial approach. Our hyperparameter study shows that there is a sweet spot for selecting the weight of hierarchy-adversarial gradients. If chosen too high, closed-set and consequently open-set performance is impaired. One approach to selecting this weight, without an open-set validation, is to select a conservatively small α that still preserves closed-set validation accuracy.

Third, existing fine-grained OSR benchmarks are based on attribute labels [53]. Our presented hierarchy-adversarial approach relies on hierarchical labels and cannot make use of attribute labels. A more general formulation of a “similarity-adversarial” approach would benefit from supporting any proxy for semantic similarity, e.g., attributes.

Finally, the benefit of pretrained models for fine-grained OSR requires further investigation. While prior work pretrained on different domains [53], pretraining on ImageNet would require the removal of 269 wild animal categories [35] that potentially overlap with iNat2021.

6. Conclusion

While previous work has mostly addressed coarser OSR we propose a new large-scale OSR benchmark which allows us to study OSR across a range of increasingly challenging semantic shifts, from coarse to fine-grained. Our results demonstrate that fine-grained OSR remains a very challenging task for current methods. We show that there are major underlying disruptive factors that hinder improved OSR performance, e.g., supervision granularity and the amount of semantic shift present in the open-set. Furthermore, hierarchical structure in the learned representation plays an important, and possibly counter-intuitive, role. We propose a hierarchy-adversarial learning strategy to discourage hierarchical structure which improves fine-grained OSR by attempting to address the “familiarity trap”. However, our analysis shows that there is still more to be done to close the gap between coarse and fine-grained OSR performance.

Acknowledgment. NL, VS, CI, SB acknowledge support by the Pioneer Centre for AI, DNRF grant number P1, and the Edinburgh–Copenhagen Strategic Partnership Award for OMA. CI is supported by “Risk-assessment of Vector-borne Diseases Based on Deep Learning and Remote Sensing” (grant number NNF21OC0069116) by the Novo Nordisk Foundation. Thanks to iNaturalist for the open data policy and to Stella Frank for valuable feedback on the manuscript.

References

[1] iNaturalist. https://www.inaturalist.org/, 2023. Accessed: 2023-11-17. 2

[2] Wentao Bao, Qi Yu, and Yu Kong. Latent space energy-based model for fine-grained open set recognition. *arXiv preprint arXiv:2309.10711*, 2023. 2

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

[4] Abhijit Bendale and Terrance E. Boult. Towards open set deep networks. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016. 2

[5] Walter Bennette, Nathaniel Hofmann, Nathaniel Wilson, and Tyler Witter. Hierarchical open-set recognition for automatic target recognition. In *2021 IEEE Symposium Series on Computational Intelligence (SSCI)*. IEEE, 2021. 3

[6] Luca Bertinetto, Romain Mueller, Konstantinos Tertikas, Sina Samangooei, and Nicholas A. Lord. Making better mistakes: Leveraging class hierarchies with deep networks. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020. 2, 3, 4, 7

[7] Julian Bitterwolf, Maximilian Mueller, and Matthias Hein. In or out? fixing imagenet out-of-distribution detection evaluation. In *International Conference on Machine Learning (ICML)*, 2023. 2, 6

[8] Guangyao Chen, Peixi Peng, Xiangqian Wang, and Yonghong Tian. Adversarial reciprocal points learning for open set recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(11), 2021. 2

[9] Elijah Cole, Kimberly Wilber, Grant Van Horn, Xuan Yang, Marco Fornoni, Pietro Perona, Serge Belongie, Andrew Howard, and Oisin Mac Aodha. On label granularity and object localization. In *European Conference on Computer Vision (ECCV)*. Springer, 2022. 3, 6, 8

[10] Elijah Cole, Xuan Yang, Kimberly Wilber, Oisin Mac Aodha, and Serge Belongie. When does contrastive visual representation learning work? In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022. 3, 6

[11] Marcos V Conde and Kerem Turgutlu. Clip-art: Contrastive pre-training for fine-grained art classification. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021. 2

[12] Yin Cui, Yang Song, Chen Sun, Andrew Howard, and Serge Belongie. Large scale fine-grained categorization and domain-specific transfer learning. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018. 1, 3

[13] Wei Dai, Wenhui Diao, Xian Sun, Yue Zhang, Liangjin Zhao, Jun Li, and Kun Fu. Camv: Class activation mapping value towards open set fine-grained recognition. *IEEE Access*, 9, 2021. 2

[14] Riccardo De Lutio, Yihang She, Stefano D'Aronco, Stefania Russo, Philipp Brun, Jan D. Wegner, and Konrad Schindler. Digital taxonomist: Identifying plant species in community scientists' photographs. *ISPRS Journal of Photogrammetry and Remote Sensing*, 182, 2021. 2, 3, 4

[15] Jia Deng, Jonathan Krause, Alexander C. Berg, and Li Fei-Fei. Hedging your bets: Optimizing accuracy-specificity trade-offs in large scale visual recognition. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, 2012. 2, 3, 4, 7

[16] Xiwen Dengxiong and Yu Kong. Ancestor search: Generalized open set recognition via hyperbolic side information learning. In *IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pages 4003–4012, 2023. 3

[17] Ankit Dhall, Anastasia Makarova, Octavian Ganea, Dario Pavllo, Michael Greeff, and Andreas Krause. Hierarchical image classification using entailment cone embeddings. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, 2020. 2, 3, 4, 7

[18] Thomas G. Dietterich and Alex Guyer. The familiarity hypothesis: Explaining the behavior of deep open set methods. *Pattern Recognition*, 132:108931, 2022. 2

[19] Andrea Frome, Greg S. Corrado, Jon Shlens, Samy Bengio, Jeff Dean, Marc' Aurelio Ranzato, and Tomas Mikolov. Devise: A deep visual-semantic embedding model. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2013. 3

[20] Yaroslav Ganin and Victor Lempitsky. Unsupervised domain adaptation by backpropagation. In *International Conference on Machine Learning (ICML)*, 2015. 2, 4, 5

[21] Yaroslav Ganin, Evgeniya Ustinova, Hana Ajakan, Pascal Germain, Hugo Larochelle, François Laviolette, Mario Marchand, and Victor Lempitsky. Domain-adversarial training of neural networks. *Journal of Machine Learning Research*, 17, 2016. 2, 5

[22] Alexander Gillert and Uwe Freiherr von Lukas. Towards combined open set recognition and out-of-distribution detection for fine-grained classification. In *International Joint Conference on Computer Vision, Imaging and Computer Graphics Theory and Applications (VISIGRAPP)*, 2021. 2

[23] Dan Hendrycks and Thomas Dietterich. Benchmarking neural network robustness to common corruptions and perturbations. *International Conference on Learning Representations (ICLR)*, 2019. 3

[24] Dan Hendrycks and Kevin Gimpel. A baseline for detecting misclassified and out-of-distribution examples in neural networks. In *International Conference on Learning Representations (ICLR)*, 2017. 3

[25] Dan Hendrycks, Steven Basart, Mantas Mazeika, Andy Zou, Joseph Kwon, Mohammadreza Mostajabi, Jacob Steinhardt, and Dawn Song. Scaling out-of-distribution detection for real-world settings. In *International Conference on Machine Learning (ICML)*, pages 8759–8773, 2022. 3

[26] 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 *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020. 3

[27] Kakani Katija, Eric Orenstein, Brian Schlining, Lonny Lundsten, Kevin Barnard, Giovanna Sainz, Oceane Boulais, Megan Cromwell, Erin Butler, Benjamin Woodward, and Katherine L. C. Bell. FathomNet: A global image database

for enabling artificial intelligence in the ocean. *Scientific Reports*, 12(1), 2022. 2

[28] Prannay Khosla, Piotr Teterwak, Chen Wang, Aaron Sarna, Yonglong Tian, Phillip Isola, Aaron Maschinot, Ce Liu, and Dilip Krishnan. Supervised contrastive learning. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2020. 3

[29] Pang Wei Koh, Shiori Sagawa, Henrik Marklund, Sang Michael Xie, Marvin Zhang, Akshay Balsubramani, Weihua Hu, Michihiro Yasunaga, Richard Lanas Phillips, Irena Gao, Tony Lee, Etienne David, Ian Stavness, Wei Guo, Berton Earnshaw, Imran Haque, Sara M Beery, Jure Leskovec, Anshul Kundaje, Emma Pierson, Sergey Levine, Chelsea Finn, and Percy Liang. Wilds: A benchmark of in-the-wild distribution shifts. In *International Conference on Machine Learning (ICML)*. PMLR, 2021. 3

[30] Jonathan Krause, Michael Stark, Jia Deng, and Li Fei-Fei. 3d object representations for fine-grained categorization. In *International Conference on Computer Vision (ICCV) Workshops*, 2013. 2

[31] Suren Kumar and Rui Zheng. Hierarchical category detector for clothing recognition from visual data. In *IEEE International Conference on Computer Vision Workshops*, 2017. 2, 3, 4

[32] Balaji Lakshminarayanan, Alexander Pritzel, and Charles Blundell. Simple and scalable predictive uncertainty estimation using deep ensembles. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2017. 3, 4, 6, 7

[33] Kibok Lee, Kimin Lee, Kyle Min, Yuting Zhang, Jinwoo Shin, and Honglak Lee. Hierarchical novelty detection for visual object recognition. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018. 3

[34] 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 (ICLR)*, 2018. 3

[35] Alexandra Sasha Luccioni and David Rolnick. Bugs in the data: How imagenet misrepresents biodiversity. In *AAAI Conference on Artificial Intelligence (AAAI)*, 2023. 8

[36] Subhransu Maji, Esa Rahtu, Juho Kannala, Matthew Blaschko, and Andrea Vedaldi. Fine-grained visual classification of aircraft. *arXiv:1306.5151*, 2013. 3

[37] Lawrence Neal, Matthew Olson, Xiaoli Fern, Weng-Keen Wong, and Fuxin Li. Open set learning with counterfactual images. In *European Conference on Computer Vision (ECCV)*, 2018. 2

[38] Mohammad Norouzi, Tomas Mikolov, Samy Bengio, Yoram Singer, Jonathon Shlens, Andrea Frome, Greg S. Corrado, and Jeffrey Dean. Zero-shot learning by convex combination of semantic embeddings. In *ICLR*, 2014. 3

[39] Eric Orenstein, Kevin Barnard, Lonny Lundsten, Geneviève Patterson, Benjamin Woodward, and Kakani Katija. The FathomNet2023 competition dataset. *arXiv preprint arXiv:2307.08781*, 2023. 2

[40] Yaniv Ovadia, Emily Fertig, Jie Ren, Zachary Nado, David Sculley, Sebastian Nowozin, Joshua Dillon, Balaji Lakshminarayanan, and Jasper Snoek. Can you trust your model's uncertainty? evaluating predictive uncertainty under dataset shift. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2019. 3, 5

[41] Lukáš Picek, Milan Šulc, Jiří Matas, and Jacob Heilmann-Clausen. Overview of FungiCLEF 2022: Fungi recognition as an open set classification problem. *In Proceedings of the Working Notes of CLEF 2022 - Conference and Labs of the Evaluation Forum.*, 2022. 2

[42] Andrés C. Rodríguez, Stefano D'Aronco, Rodrigo Caye Daudt, Jan D. Wegner, and Konrad Schindler. Zero-shot bird species recognition by learning from field guides. In *IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, 2024. 1, 3, 5

[43] Abhijit Guha Roy, Jie Ren, Shekoofeh Azizi, Aaron Loh, Vivek Natarajan, Basil Mustafa, Nick Pawlowski, Jan Freyberg, Yuan Liu, Zach Beaver, Nam Vo, Peggy Bui, Samantha Winter, Patricia MacWilliams, Greg S. Corrado, Umesh Telang, Yun Liu, Taylan Cemgil, Alan Karthikesalingam, Balaji Lakshminarayanan, and Jim Winkens. Does your dermatology classifier know what it doesn't know? detecting the long-tail of unseen conditions. *Medical Image Analysis*, 75, 2022. 2

[44] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, Alexander C. Berg, and Li Fei-Fei. Imagenet large scale visual recognition challenge. *International Journal of Computer Vision*, 115, 2015. 2, 3

[45] Walter J. Scheirer, Anderson de Rezende Rocha, Archana Sapkota, and Terrance E. Boult. Toward open set recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 35(7), 2012. 1, 2

[46] Terence Sim, Simon Baker, and Maan Bsat. The CMU pose, illumination, and expression (PIE) database. *IEEE International Conference on Automatic Face Gesture Recognition*, 2002. 3

[47] Paolo Simeone, Raúl Santos-Rodríguez, Matt McVicar, Jefrey Lijffijt, and Tijl De Bie. Hierarchical novelty detection. In *Advances in Intelligent Data Analysis*. Springer, 2017. 3

[48] Jiayin Sun, Hong Wang, and Qiulei Dong. Hierarchical attention network for open-set fine-grained image recognition. *IEEE Transactions on Circuits and Systems for Video Technology*, 2023. 2

[49] Yiyou Sun, Yifei Ming, Xiaojin Zhu, and Yixuan Li. Out-of-distribution detection with deep nearest neighbors. In *International Conference on Machine Learning (ICML)*. PMLR, 2022. 3, 4, 6, 7

[50] Mehmet Ozgur Turkoglu, Stefano D'Aronco, Gregor Perich, Frank Liebisch, Constantin Streit, Konrad Schindler, and Jan Dirk Wegner. Crop mapping from image time series: Deep learning with multi-scale label hierarchies. *Remote Sensing of Environment*, 264, 2021. 2, 3, 4

[51] 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 *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018. 2, 3, 6

[52] Grant Van Horn, Elijah Cole, Sara Beery, Kimberly Wilber, Serge Belongie, and Oisin Mac Aodha. Benchmarking

representation learning for natural world image collections. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021. 2, 5

[53] Sagar Vaze, Kai Han, Andrea Vedaldi, and Andrew Zisserman. Open-set recognition: A good closed-set classifier is all you need. In *International Conference on Learning Representations (ICLR)*, 2022. 1, 2, 3, 4, 5, 6, 7, 8

[54] Nakul Verma, Dhruv Mahajan, Sundararajan Sellamanickam, and Vinod Nair. Learning hierarchical similarity metrics. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, 2012. 3

[55] Catherine Wah, Steve Branson, Peter Welinder, Pietro Perona, and Serge Belongie. The Caltech-UCSD Birds-200-2011 dataset. 2011. 2

[56] Dequan Wang, Zhiqiang Shen, Jie Shao, Wei Zhang, Xiangyang Xue, and Zheng Zhang. Multiple granularity descriptors for fine-grained categorization. In *International Conference on Computer Vision (ICCV)*, 2015. 3

[57] Xiu-Shen Wei, Yi-Zhe Song, Oisin Mac Aodha, Jianxin Wu, Yuxin Peng, Jinhui Tang, Jian Yang, and Serge Belongie. Fine-grained image analysis with deep learning: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(12):8927–8948, 2022. 3

[58] Hui Wu, Michele Merler, Rosario Uceda-Sosa, and John R. Smith. Learning to make better mistakes: Semantics-aware visual food recognition. In *ACM International Conference on Multimedia*, 2016. 2, 3, 4

[59] Jingkang Yang, Kaiyang Zhou, Yixuan Li, and Ziwei Liu. Generalized out-of-distribution detection: A survey. *arXiv preprint arXiv:2110.11334*, 2021. 3

[60] Hongjie Zhang, Ang Li, Jie Guo, and Yanwen Guo. Hybrid models for open set recognition. In *European Conference on Computer Vision (ECCV)*. Springer, 2020. 2

[61] Richard Zhang, Phillip Isola, Alexei A. Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018. 3

[62] Bingchen Zhao, Shaozuo Yu, Wufei Ma, Mingxin Yu, Shenxiao Mei, Angtian Wang, Ju He, Alan Yuille, and Adam Kortylewski. OOD-CV: a benchmark for robustness to out-of-distribution shifts of individual nuisances in natural images. In *European Conference on Computer Vision (ECCV)*. Springer, 2022. 3