Open-Insect: Benchmarking Open-Set Recognition of Novel Species in Biodiversity Monitoring

Yuyan Chen[1,2], **Nico Lang**[3], **B. Christian Schmidt**[4], **Aditya Jain**[2],
Yves Basset[5,6,7], **Sara Beery**[8], **Maxim Larrivée**[9], **David Rolnick**[1,2],
[1]McGill University [2]Mila - Quebec Artificial Intelligence Institute [3]University of Copenhagen
[4]Agriculture and Agri-food Canada [5]Smithsonian Tropical Research Institute
[6]Biology Center, Czech Academy of Sciences [7]Maestria de Entomologia, University of Panama
[8]Massachusetts Institute of Technology [9]Montréal Insectarium
`https://yuyan-c.github.io/open-insect-project/`

Abstract

Global biodiversity is declining at an unprecedented rate, yet little information is known about most species and how their populations are changing. Indeed, some 90% of Earth's species are estimated to be completely unknown. Machine learning has recently emerged as a promising tool to facilitate long-term, large-scale biodiversity monitoring, including algorithms for fine-grained classification of species from images. However, such algorithms typically are *not* designed to detect examples from categories unseen during training – the problem of open-set recognition (OSR) – limiting their applicability for highly diverse, poorly studied taxa such as insects. To address this gap, we introduce Open-Insect, a large-scale, fine-grained dataset to evaluate unknown species detection across different geographic regions with varying difficulty. We benchmark 38 OSR algorithms across three categories: post-hoc, training-time regularization, and training with auxiliary data, finding that simple post-hoc approaches remain a strong baseline. We also demonstrate how to leverage auxiliary data to improve species discovery in regions with limited data. Our results provide insights to guide the development of computer vision methods for biodiversity monitoring and species discovery.

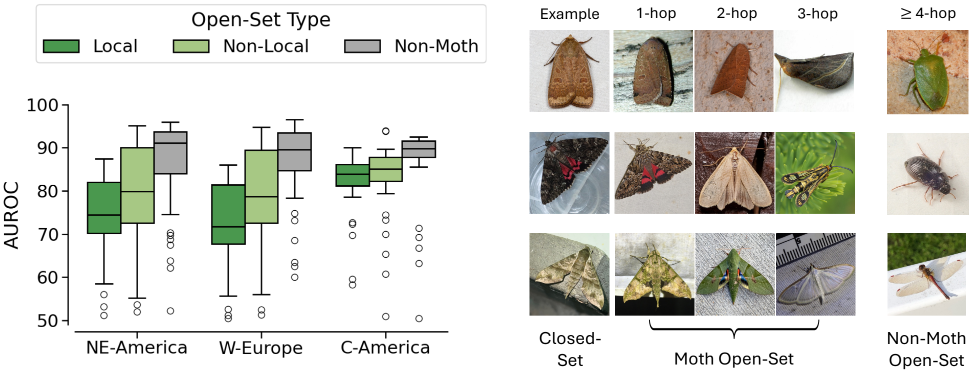


Figure 1: **Open-Insect benchmark results on three geographical regions with varying difficulty.** The Open-Insect benchmark includes images of thousands of highly visually similar moth species, along with non-moth arthropods, divided by geographic region. **Left**: Results from 38 OSR methods on three open-set types i) *Local* moth, ii) *Non-local* moth, and iii) *Non-moth* (see Table 2). **Right**: Visual dissimilarity across taxonomic levels: 1-hop (same genus), 2-hop (different genus, same family), 3-hop (different family within *Lepidoptera*), and non-moths (different order, ≥4 hops).

39th Conference on Neural Information Processing Systems (NeurIPS 2025) Track on Datasets and Benchmarks.

1 Introduction

The use of machine learning (ML) for species recognition can greatly facilitate long-term, large-scale biodiversity conservation [80]. In recent years, significant progress has been made in developing sophisticated ML and computer vision tools in biodiversity, including a variety of challenging benchmarks [82, 81, 85, 10] and foundation models [23, 24, 74, 91], which span hundreds of millions of images and hundreds of thousands of species.

Most ML models are trained with the closed-world assumption, assuming that at inference time, the model will only encounter categories that are seen during training. However, in biodiversity monitoring, the closed-world assumption is often violated for several reasons. Firstly, an estimated 86 percent of terrestrial species and 91 percent of marine species remain undescribed [59]. Regardless of the size and diversity of the training dataset, unknown species can readily occur when deploying a model in practice, especially with less well-studied groups such as insects or deep-sea fauna. Secondly, not all described species have data available for training. For example, there are over 2 million species according to the latest Catalogue of Life checklist [9], but only 65 percent of them have a record on the Global Biodiversity Information Facility (GBIF)[1], which compiles records from across biodiversity platforms such as iNaturalist[2]. Thirdly, very often, species recognition models are trained with regional checklists of species that are of interest in specific use cases. Such models cannot correctly identify species that are out of scope even if they are well documented.

For these reasons, open-set recognition (OSR), which aims at accurately classifying closed-set samples (samples that belong to a category seen during training) and detecting open-set (OS) samples, should be a crucial component of ML for biodiversity monitoring. Beyond improving the accuracy of existing identification systems, OSR algorithms can help to pinpoint previously undiscovered species and invasive or introduced species in specific geographic regions. Most existing OSR benchmarks [95, 22, 34, 7] are derived from ImageNet-1K [67] and ImageNet-21K [13], which do not capture the complexities and fine-grained categories of biodiversity data [56] or its long-tailed distribution. While some OSR benchmarks do contain biodiversity categories, they are relatively small in scale and focus on well-studied taxa such as birds [84] and other vertebrates [37]. As a result, performance on these benchmarks may not accurately represent performance in new species detection for poorly documented, highly diverse taxa.

An ideal benchmark dataset for new species detection should contain taxa with plentiful undiscovered species. Since about two-thirds of all animals are insects [9] and over 80 percent of insect species remain undescribed [75], we introduce Open-Insect, a large-scale, fine-grained image dataset which focuses on insects (see Fig. 1 for examples). Open-Insect consists of closed-, open-set, and auxiliary splits for three geographical regions: Northeastern North America, Western Europe, and Central America. Following Tobler's first law of geography[3], we hypothesize that the difficulty of detecting semantic shifts correlates with geographic proximity. Therefore, Open-Insect utilizes geographical metadata to study local and non-local semantic shifts. Local open-set species may not only be harder to detect, but also reflect challenges encountered when deploying species-recognition models in practice. For each region, we also include a realistic auxiliary dataset for OSR methods that benefit from training with such data.

Our contributions can be summarized as follows:

1. We introduce Open-Insect, a large-scale, extremely fine-grained biodiversity dataset focused on insects for open-set recognition. Open-Insect allows us to benchmark the performance of 36 existing OSR methods for species discovery and invasive species detection.
2. We show that the quality of auxiliary data is crucial to improve OSR performance, and present a simple approach to benefit from auxiliary training data.
3. We observe that the maximum softmax probability (MSP) and methods derived from it remain a strong baseline for fine-grained OSR.

---

[1] https://www.gbif.org/ (accessed 2025-01-29)

[2] https://www.inaturalist.org/ (accessed 2025-01-29)

[3] "*Everything is related to everything else, but near things are more related than distant things.*"

2 Related Work

2.1 OSR and OOD Detection

The goal of OSR is to accurately classify the closed-set categories and detect the open-set ones [69]. OpenMax [6] is the first deep learning-based method for open-set recognition. More recent methods include ARPL [11], OpenGAN [45], and hierarchy-adversarial learning [48]. Out-of-distribution (OOD) detection is a broader term that refers to any setting in which one must detect test samples drawn from a different distribution than the training distribution. The shift can be either covariate or semantic. Covariate shift occurs when OOD samples come from a different input space, while semantic shift arises when new categories (labels) occur in the test set [94]. The majority of recent OOD detection methods focus on semantic shift and tackle the same problem as OSR [94], making them suitable for species discovery as well. Hence, we also compare against these OOD detection methods in our work.

We categorize the aforementioned OOD detection methods into 1) post-hoc, 2) training-time regularization, and 3) training with auxiliary data. *Post-hoc* methods, as the name suggests, do not require any further training and can be easily integrated into an existing pipeline. These methods utilize the outputs of an already trained classifier and develop a score function to map the outputs to a real number [6, 31, 26, 50, 49, 49, 64, 68, 54, 45, 39, 76, 29, 84, 30, 87, 78, 77, 73, 17, 101, 3, 51, 43, 52]. These outputs can be, for instance, features [87, 64, 49], logits [29], gradients [39], or softmax probability [31]. For methods that require further training, a number of methods do not require extra training data, but improve OSR performance by *regularization* with new loss functions or model architectures, such as [15, 89, 11, 38, 33, 58, 40, 20, 45]. However, OSR can benefit from extra data if they are available. Early works that explored the usefulness of auxiliary data include Entropic Open-Set Loss [16] and Outlier Exposure [32] . Recently, an increasing number of methods have been developed which involve *training models with auxiliary data* [36, 57, 96, 71, 90, 21, 19, 42, 60, 4, 5], utilizing either labeled or unlabeled data.

2.2 Benchmarks and Datasets

OSR tasks usually involve splitting categories of a dataset into the closed and the open set. On the other hand, OOD detection tasks typically involve taking an entire dataset, such as ImageNet [67], as in-distribution (ID) and several other datasets as OOD, such as Texture [47] or iNaturalist [82]. OpenOOD [93] is the first benchmark to standardize the evaluation of OOD detection and OSR methods. The 4 OOD benchmarks follow the conventional setup as mentioned above. The 4 OSR benchmarks were constructed by splitting small-scale datasets, namely MNIST [14], CIFAR10 [46], CIFAR100 [46], and TinyImageNet-200 [79] into the closed and the open sets. A series of datasets has also been developed in the same way by splitting ImageNet-1K [67] and ImageNet-21K [13] to evaluate semantic shift detection while minimizing covariate shifts, including ImageNet-OOD [95], COOD [22], ImageNet-O [34], and OpenImageNet [87]. OpenOOD v1.0 [93] and v1.5 [100] are the only benchmarks, to our knowledge, to include an auxiliary dataset, TinyImageNet [79], for the 4 OOD datasets. In many works [36, 57, 5, 96, 71, 90, 21, 19, 42, 60, 4], training with auxiliary data methods are evaluated with different ID, OOD, and auxiliary data splits, making a fair comparison of these works challenging.

The Semantic Shift Benchmark (SSB) [84], Combined Out-of-Distribution Detection (COOD) [37], and iNat21-OSR [48] include OSR or OOD detection datasets that consist of biodiversity categories. SSB consists of an OSR dataset based on Caltech-UCSD Birds-200-2011(CUB) [86]. COOD [37] used Norwegian vertebrates [70] as the ID set and Non-Norwegian vertebrates [70] as the OOD set to evaluate OSR in hierarchical classification. Both SSB and COOD are relatively small-scale and focus on well-studied groups of animals. iNat21-OSR [48] was developed to compare OSR performance at different semantic distances from coarse to fine-grained in the taxonomic hierarchy, from the same genus to different kingdoms. None of these datasets includes auxiliary training data.

2.3 ML for Biodiversity Monitoring

Machine learning is increasingly used in species recognition to support biodiversity monitoring. A number of large-scale, fine-grained species recognition datasets have been introduced, such as iNat18 [82], iNat21 [81], Tree-of-Life 10M [74], and BioTrove [91]. Recently, foundation models such as

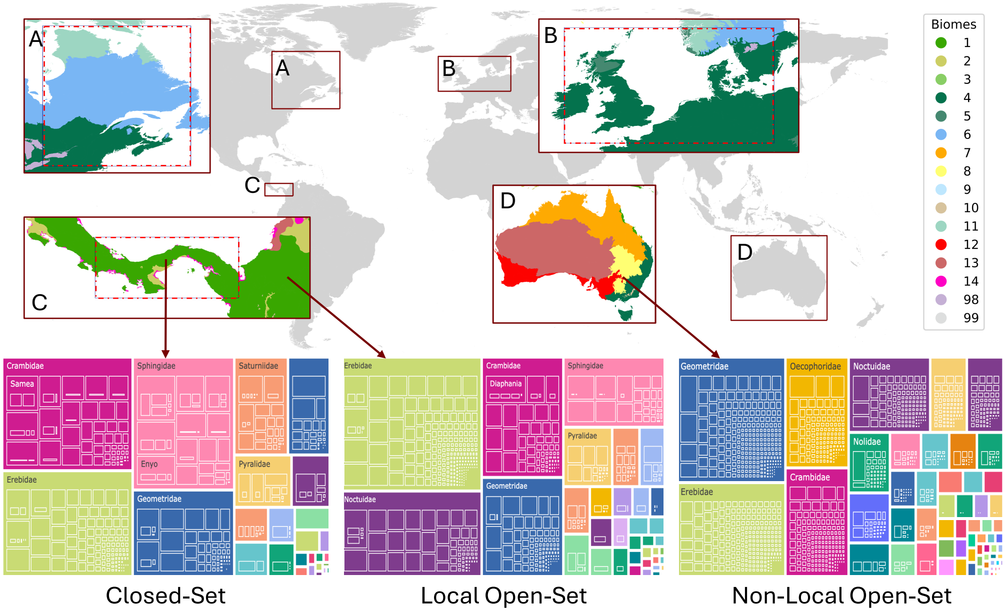


Figure 2: **Open-Insect dataset overview.** Regions A, B, and C correspond to closed-sets and local open-sets, while region D corresponds to the non-local open-set. (Region A: NE-America; B: W-Europe; C: C-America; D: Australia.) The tree maps visualize the taxonomic distribution of moth families in regions C and D, where nested boxes denote genera and species, and box size indicates the relative number of images. The same family is *colored consistently* across the three treemaps. Local open-set species are more similar to the closed-set than non-local ones. Biome codes and tree maps for the other two regions are provided in the Appendix.

BioCLIP [74] and BioTrove-CLIP [91] have been developed to achieve generalization across datasets, domains, and tasks, with insect-specific models like CLIBD [25], trained on BIOSCAN-1M [23] and BIOSCAN-5M [24], and Insect-Foundation [61], trained on Insect-1M, marking significant progress in this area. A detailed comparison between the insect datasets and Open-Insect is provided in the Appendix. Some of these models achieve zero-shot classification of unseen species, by leveraging prior knowledge of these species such as scientific names, reference images, or DNA barcodes. This technique is not applicable to flagging undescribed species, which lack such prior information, but can be useful for identifying rare species or supporting checklists of species of interest.

3 The Open-Insect Dataset

Of all insects, we focus on moths in this work, as moths are easily attracted to ultraviolet camera traps [8] and many can be visually identified to the species level, allowing automatic large-scale long-term monitoring [66, 83, 41]. These monitoring systems are expected to encounter a plethora of novelties [65, 72], making it especially timely to develop ML-based models that can accommodate undocumented species and accelerate species discovery. We curate Open-Insect based on the AMI dataset [41], a large-scale, fine-grained dataset consisting of 5,364 moth species and 12 groups of non-moth arthropods in three regions: Northeastern North America (NE-America), Western Europe (W-Europe), and Central America (C-America) (see Fig. 2: areas inside the dashed bounding boxes A, B, and C). We summarize the Open-Insect data distribution in Table 1.

3.1 The Closed-Set

Our benchmark dataset uses the AMI-GBIF dataset [41] as in-distribution data. We follow the regional splits in the AMI dataset, as the first two regions represent well-documented regions, while the last represents regions with very high biodiversity but limited data. Beside the AMI-GBIF dataset, we compiled 2,912,168 images of 28,388 moth species from GBIF to minimize covariate shift for open-set and auxiliary species (see the Appendix for curation details).

3.2 The Open-Sets

We designed three open-sets: 1) Local moth (O-L), 2) Non-Local moth (O-NL), 3) and Non-Moth (O-NM), to simulate different kinds of unseen species that a model may encounter in the wild – respectively: 1) species related to those known by the model but which are undocumented (no data or new to science), 2) introduced or invasive species that have only recently colonized the region, and 3) species that are not the focus of the model, but may occur in the monitoring system.

Local moths. Since we have very limited data on truly new species, we need to use described species to simulate them. To gather a set of local open-set species, we extended each region by 1 degree in latitude and 3 degrees in longitude, and then compiled a checklist containing species that have at least one occurrence in the extended area (Fig. 2: areas inside the solid bounding boxes A, B, and C) and excluded all species already present in the closed-set. Since all three regions share similar biomes with their neighboring areas (see Fig. 2), it is likely that many of these local open-set species are also present but yet to be observed in the given region, making them ideal substitutes for new species.

Figure 3: **Examples of likely novel species from BCI.** These species show greater than 7% DNA barcode divergence from their closest match in the BOLD database, well above the 1.5% species-level cutoff.

For C-America, we also include 197 images recently collected from Barro Colorado Island (BCI), Panama, gathered by our team to provide a maximally realistic test case. This dataset split, referred to as **O-B** in Table 1, includes 133 open-set species—59 of which are likely undescribed by science (see examples in Fig. 3). Though these species have not gone through the formal accreditation process yet, they exhibit over 1.5% DNA sequence divergence from their closest match in the Barcode of Life Data (BOLD) database [62]. This is a commonly used threshold for distinguishing similar but distinct sibling species [98, 63]. The O-B split was not used for the benchmark results reported in Table 2, but to assess 1) whether methods that perform well on Open-Insect also generalize well to potentially new species, and 2) whether the OSR score threshold, determined by fixing the false positive rate on Open-Insect, remains effective for species discovery in real-world scenarios.

Non-local moths. Here we simulated non-native species (such as introduced invasive species) by selecting all species occurring in Australia (Fig. 2: region D) and falling outside the closed-sets and local open-sets. Australia was picked since it is geographically isolated and far from all closed-set regions. This set contains 4,662 species and 95,597 samples. In Fig. 1, we show the visual dissimilarity of moth species at different taxonomic distances: 1-hop (from the same genus), 2-hop (from a different genus but same family), and 3-hop (from a different family). The detailed comparison of the local and non-local open-sets are summarized in the Appendix. In short, local open-sets have substantially more 1-hop species than non-local.

Non-moths. To evaluate ML algorithms' performance on rejecting species that are beyond the scope of study, but may occur in monitoring sites, we randomly sampled 35,000 images of non-moth species from the AMI-GBIF dataset. These species are mostly other insects, along with some arachnids. Examples are shown in Fig. 1.

3.3 Auxiliary Data

As mentioned in Sec. 2.2, when comparing OSR algorithms that use auxiliary data, it is important that such data be consistent across methods. This motivates us to explicitly include an auxiliary training set for each region in the Open-Insect benchmark, in contrast to previous benchmarks. By including auxiliary datasets, our goal is not merely to expand the training set, but to curate a *unified* and *comprehensive* benchmark to facilitate the development and evaluation of OSR methods that were not previously evaluated with biodiversity-related datasets [84, 37, 48].

For each regional dataset, we constructed the auxiliary set (AUX) as follows. We first excluded all species that are in the closed-set, the local open-set, or the non-local open-set. Since species within the same genus can be highly similar, we also excluded species from any genus present in the open-set but not the closed-set, since it is possible that a model could learn to simply declare all species from this genus as unknown, thereby inflating performance metrics. For NE America and W Europe, we selected 8,000 species for training, using 20 images per species. For Central America, we selected 4,000 species, each with 20 images. We also randomly sampled a small set of images from the auxiliary set to construct regional validation sets for hyperparameter tuning of the post-hoc methods. Details can be found in Table 1.

4 Methods

Existing methods. Similar to OpenOOD v1.5 [100], we divided the methods into three categories: 1) post-hoc; 2) training-time regularization; and 3) training with auxiliary data. For post-hoc methods, we evaluate all such methods in OpenOOD v1.5 [100] and also more recent methods, including NECO [3], FDBD [51], RP [43], and NCI [52]. For training-time regularization, we evaluate ConfBranch [15], G-ODIN [38], ARPL [11], LogitNorm [89], and RotPred [33]. Though OpenGAN [45] is considered as post-hoc in OpenOOD v1.5 [100], we followed OpenOOD [93] and categorize it as "training-time regularization" because it still requires training the generator and the discriminator with features extracted by the classifier, making it more complicated to use than other post-hoc methods. We evaluated all "training with extra data" methods in OpenOOD v1.5 [100] except for MCD [97], due to its extremely low closed-set classification accuracy, and included Energy [54].

Table 1: **The Open-Insect dataset.** For each region, we consider species in AMI-GBIF as the closed-set, and species from nearby regions as local open-set. We also provide auxiliary data with other species that are available at train time. In addition, the Open-Insect dataset provides a non-local open-set (O-NL) as well as a non-moth open-set (O-NM). Species in the closed-, O-L, or AUX sets may overlap across the three regions, but not within a region. These images were compiled from public data on GBIF. For C-America, we also include 197 images of 57 potentially new species (denoted as "O-B") collected by our team in BCI.

Region	**Type**	**Moth species**	**Train images**	**Val images**	**Test images**
NE-A	Closed	2,497	870,336	102,987	206,620
	O-L	617	-	-	113,634
	AUX	8,000	160,000	19,439	-
W-E	Closed	2,603	1,177,125	134,095	268,113
	O-L	458	-	-	78,648
	AUX	8,000	160,000	21,868	-
C-A	Closed	636	72,188	9,054	18,163
	O-L	1,724	-	-	111,879
	O-B	133	-	-	197
	AUX	4,000	80,000	9,054	-
-	O-NL	4,662	-	-	95,597
	O-NM	-	-	-	35,000

Proposed baselines for training with auxiliary data. We additionally propose two simple baselines that utilize auxiliary data. Let C be the number of species in the closed-set and A the number of species in the auxiliary set. We preserve the original C-dimensional classification head and simultaneously train a second linear classification head with either $C+1$ or $C+A$ dimensions – i.e., treating the auxiliary set either as one new class or as A new classes. We call these two methods *NovelBranch* and *Extended*, respectively. In our experiments, the models were optimized with the cross-entropy of both classification problems sharing the feature extractor. At test time, the OSR scores were computed from the closed-set head so that ID accuracy is preserved.

Implementation details. The implementation follows OpenOOD[4] [93]. We used a ResNet-50 backbone [27] for all methods, following a standard practice in the OSR [48], OOD detection [93, 100], and species recognition literature [81]. We used the AdamW optimizer [55] and applied cosine scheduling [28] to the learning rate. The weight decay was constant (10^{-4}) during training. All basic classifiers were trained from scratch for 120 epochs, with 0.01 initial learning rate (lr) and 6 warm-up epochs. For all training methods (with or without auxiliary data), we fine-tuned species-level classifiers for 30 additional epochs, with 0.001 initial lr and 2 warm-up epochs. We trained three basic classifiers per region and report the mean and standard deviation of their performance in Table 2. The models for training methods were fine-tuned using one randomly selected basic classifier. More training details can be found in the Appendix.

Using pre-trained weights. It has been shown that ImageNet pre-trained weights can improve accuracy on species recognition when the training set is small [81]. Since the C-America closed-set is much smaller than those for the other two regions, we trained a ResNet50 classifier using the ImageNet-1K pretrained weights to see if we can further improve the OSR performance. No moths are present in ImageNet-1K, so using the pre-trained weights should not affect evaluation on local

---

[4] https://github.com/Jingkang50/OpenOOD/tree/main (accessed 2025-01-29)

Table 2: **Benchmarking results on Open-Insect.** We evaluate approaches falling into three categories: i) post-hoc methods, ii) training-time regularization, and iii) training with auxiliary data. Results are shown for the three regions in Open-Insect: NE-America, W-Europe, C-America. For each of the three open-set splits – local (L), non-local (NL), and non-moth (NM) – the AUROC is reported along with the accuracy of the closed-set. The best result within each category is **bold**, and the overall best result is **bold and underlined**. For post-hoc methods, we report the mean$_{\text{(standard deviation)}}$ from three training runs.

	NE-America				W-Europe				C-America			
	L	NL	NM	Acc.	L	NL	NM	Acc.	L	NL	NM	Acc.
Post-hoc methods												
OpenMax [6]	70.0 (1.3)	75.9 (1.8)	63.8 (3.8)	89.7 (0.1)	68.9 (0.4)	75.9 (0.8)	60.0 (1.6)	88.4 (0.4)	81.9 (0.3)	82.7 (0.3)	86.6 (0.3)	85.4 (0.4)
MSP [31]	86.7 (0.3)	93.6 (0.2)	94.3 (0.3)		**86.0** (0.7)	93.2 (0.6)	94.5 (0.0)		86.7 (0.2)	88.0 (0.2)	89.0 (0.4)	
TempScale [26]	**86.8** (0.3)	**93.7** (0.2)	94.1 (0.3)		**86.0**(0.6)	**93.5** (0.5)	**94.5** (0.1)		85.0 (1.76)	86.5 (1.45)	89.3 (0.62)	
ODIN [50]	84.7 (0.3)	91.5 (0.6)	89.2 (0.5)		81.4 (3.2)	89.5 (2.7)	86.1 (3.6)		83.5 (0.3)	84.8 (0.3)	85.6 (0.3)	
MDS [49]	73.6 (1.4)	79.2 (1.7)	**94.4** (0.7)		71.7 (1.4)	77.1 (2.0)	94.1 (1.4)		82.1 (0.3)	83.5 (0.4)	92.1 (0.4)	
MDSEns [49]	53.2 (0.2)	53.6 (0.2)	62.2 (0.7)		50.5 (0.0)	56.0 (0.2)	62.5 (0.4)		58.2 (0.1)	60.7 (0.4)	69.3 (0.6)	
RMDS [64]	79.9 (1.5)	87.9 (1.3)	90.4 (3.2)		79.2 (1.9)	88.2 (1.4)	90.5 (4.9)		85.0 (0.3)	87.1 (0.2)	91.9 (0.4)	
Gram [68]	-	-	-		-	-	-		46.1 (0.0)	50.9 (0.0)	50.5 (0.0)	
EBO [54]	72.8 (3.0)	78.2 (4.0)	91.0 (3.6)		69.7 (3.2)	76.0 (3.4)	88.8 (6.7)		85.3 (0.3)	86.7 (0.4)	91.8 (0.3)	
GradNorm [39]	49.3 (2.5)	52.0 (3.3)	22.9 (2.7)		51.1 (0.5)	52.5 (1.2)	25.1 (8.2)		33.0 (1.0)	32.1 (1.0)	18.3 (0.5)	
ReAct [76]	73.1 (2.9)	79.9 (3.8)	91.1 (2.1)		70.0 (2.6)	78.7 (2.6)	89.6 (5.2)		83.6 (0.4)	85.1 (0.5)	91.4 (0.4)	
MLS [29]	73.8 (3.1)	79.4 (4.1)	91.4 (3.5)		70.7 (3.4)	77.2 (3.6)	89.3 (6.5)		86.5 (0.3)	87.9 (0.3)	91.6 (0.4)	
KLM [30]	-	-	-		-	-	-		83.5 (0.3)	84.7 (0.4)	88.3 (0.3)	
VIM [87]	72.5 (4.5)	78.0 (5.7)	83.4 (10.2)		68.6 (5.3)	76.3 (5.1)	80.5 (14.8)		83.3 (0.3)	84.8 (0.3)	**92.5**(0.4)	
KNN [78]	64.3 (17.6)	69.0 (21.7)	70.3 (35.3)		67.3 (11.4)	71.1 (19.6)	74.8 (28.0)		80.9 (0.2)	81.8 (0.3)	88.3 (0.3)	
DICE [77]	59.2 (11.6)	62.9 (15.4)	74.6 (26.4)		57.3 (10.9)	62.3 (15.3)	78.4 (21.3)		23.2 (0.5)	21.5 (0.6)	13.8 (0.7)	
RankFeat [73]	61.3 (3.6)	64.8 (4.8)	69.7 (13.7)		52.6 (2.4)	51.3 (6.5)	63.5 (8.2)		72.3 (1.1)	73.3 (1.5)	86.4 (1.2)	
ASH [17]	72.5 (2.9)	78.0 (3.7)	91.5 (3.4)		69.2 (3.2)	75.9 (2.9)	89.1 (6.6)		82.9 (0.5)	84.5 (0.5)	92.0 (0.3)	
SHE [101]	58.5 (14.4)	62.9 (18.4)	67.5 (30.2)		58.4 (9.9)	61.2 (15.6)	68.5 (24.3)		78.6 (0.4)	79.4 (0.8)	87.7 (0.2)	
NECO [3]	51.2 (2.6)	55.2 (4.3)	52.2 (3.7)		-	-	-		69.7 (0.3)	70.0 (0.2)	66.8 (0.6)	
FDBD [51]	80.5 (3.2)	88.1 (3.9)	93.8 (2.2)		78.2 (5.1)	86.1 (4.9)	91.3 (5.0)		86.0 (0.3)	87.8 (0.2)	91.6 (0.4)	
RP$_{\text{MSP}}$ [43]	86.7 (0.3)	93.6 (0.2)	94.3 (0.3)		**86.0** (0.7)	93.2 (0.6)	**94.5** (0.0)		**86.8** (0.2)	**88.1** (0.2)	89.1 (0.4)	
RP$_{\text{ODIN}}$ [43]	81.7 (1.1)	91.5 (0.8)	90.7 (0.9)		82.1 (1.4)	91.4 (1.3)	92.3 (1.5)		84.2 (0.2)	86.5 (0.2)	89.1 (0.2)	
RP$_{\text{EBO}}$ [43]	70.8 (3.6)	74.6 (4.9)	89.3 (4.5)		67.7 (3.4)	72.5 (3.8)	86.7 (8.0)		85.7 (0.3)	87.6 (0.4)	**92.5** (0.3)	
RP$_{\text{GradNorm}}$ [43]	39.6 (2.9)	38.4 (4.1)	17.9 (4.3)		43.6 (1.3)	41.4 (1.0)	21.4 (8.2)		21.4 (0.5)	20.0 (0.5)	12.7 (0.4)	
NCI [52]	72.1 (5.3)	78.4 (5.5)	91.6 (3.4)		70.9 (2.9)	77.4 (3.2)	89.8 (6.0)		85.1 (0.3)	87.0 (0.3)	92.4 (0.4)	
Training regularization												
ConfBranch [15]	67.2	70.5	93.8	89.8	61.4	59.2	90.8	86.0	79.8	81.9	91.0	85.5
OpenGAN [45]	41.1	41.7	12.7	88.6	52.7	56.9	28.2	87.9	46.1	48.6	40.6	82.2
LogitNorm [89]	80.6	87.3	95.3	85.5	80.8	87.7	95.6	84.5	**87.6**	**89.5**	90.5	**85.7**
ARPL [11]	**82.1**	87.5	93.6	89.7	**81.4**	86.2	93.2	88.6	85.8	87.3	89.8	85.1
G-ODIN [38]	80.7	**88.7**	93.6	**90.0**	80.6	**89.2**	**96.5**	**88.7**	72.7	74.4	71.4	83.1
RotPred [33]	77.6	85.7	**95.9**	89.7	73.9	85.9	96.2	88.6	82.2	84.9	**91.3**	85.3
Training with auxiliary data												
OE [32]	79.8	86.3	90.1	85.3	75.4	84.6	89.6	83.6	89.5	**94.0**	**92.1**	84.0
UDG [92]	75.1	81.9	91.5	80.8	-	-	-	-	80.8	83.1	87.9	76.0
MixOE [99]	86.2	92.5	**94.4**	**90.3**	**85.2**	91.9	**94.1**	**89.1**	86.2	87.9	90.1	84.9
Energy [54]	**87.4**	**95.1**	92.5	89.6	84.6	**94.8**	89.5	88.4	**90.0**	93.8	91.2	84.2
NovelBranch (Ours)	85.5	94.1	90.0	89.9	83.9	93.8	91.7	88.6	87.8	89.7	91.1	85.6
Extended (Ours)	83.5	92.2	86.1	89.8	82.6	91.9	89.4	88.6	86.9	89.0	89.3	**85.8**

or non-local open-sets. Additionally, we finetuned two ViT-B-16 models [18] - one trained with ImageNet-1K and the other is the visual encoder of BioCLIP [74], a foundation model for species recognition which is increasingly used for biodiversity-related tasks. We evaluated the performance of the two models on a subset of the open-set, excluding moths that are in its training set. Details of this subset are presented in the Appendix.

5 Results

We present the performance of 38 OSR methods in Table 2. Following OpenOOD v1.5 [100], we consider open-set samples as positive and closed-set samples as negative, and use the area under the ROC curve (AUROC) as the evaluation metric for OSR. We report the classification accuracy (Acc.) for the closed-set and AUROC for OSR for each method on the Open-Insect benchmark. Some of the methods listed in Table 2 under "Training regularization" and "Training with auxiliary data" require combination with a post-hoc method of the user's choice. In these cases, we tested 4 simple yet effective post-hoc methods: MSP[31], EBO [54], MLS [29], and TempScale [26]. Results in Table 2 show the best performing method of each. Full results are listed in the Appendix. We also report more recently proposed OSR metrics, including OSCR [16], AUOSCR [84], OpenAUC [88], and OOSA [12] in the Appendix. Some results are shown as "-" either due to extremely long inference time (KLM) or very large memory requirement (Gram, NECO and UDG).

Difficulty of the three open-set types. Fig. 1 summarizes results for each of the three types of open-sets in each of the three regions. The maxima shown in the box plots show a clear trend that local open-set species are substantially more difficult to detect than non-local and non-moth open-set species. This validates our hypothesis that novel species from the same geographical region are more challenging to detect. One explanation for this is that open-set species that are geographically closer are also, on average, taxonomically closer to the species seen during training (see Table 6). Taxonomically closer open-set species are harder to detect as they tend to exhibit more visual similarity with closed-set species [84, 48]. However, for all regions, the best performing method can achieve over 85% AUROC, indicating strong potential for automating species discovery with OSR in computer vision.

Effect of pre-trained weights. In general, we observe increased AUROC and closed-set accuracy when we utilize the ImageNet-1K pre-trained weights to initialize the weights of the base classifier. All of them outperformed their training-from-scratch counterpart. Our results show that for small-scale dataset, leveraging pre-trained weights can be helpful for OSR, even if neither the closed-set nor the open-set categories are included in the pre-training dataset. Our proposed methods achieve the two highest AUROC on local open-set, 91.19 (Extended) and 92.05 (NovelBranch) and LogitNorm has the best performance of non-local open-set, with the AUROC increasing from 94.04 to 95.38 (see the Appendix for the full result). When comparing the BioCLIP weights with the ImageNet-1K weights of ViT-B-16, we find that BioCLIP weights give better performance on a subset of Open-Insect, where open-set species present in BioCLIP's training set are excluded (see Table 3).

5.1 Post-hoc Methods

Though no method consistently outperforms the others across datasets, we observe that maximum softmax probability (MSP) [31] and MSP-based methods, namely TempScale [26] and RP_{MSP} [43], remain a strong baseline. For open-set moth species (both local and non-local), TempScale performs the best for NE-America and W-Europe and RP_{MSP} performs the best for C-America.

Table 3: **Comparison of pretrained weights.** OSR performance of models finetuned from BioCLIP or ImageNet-1K.

Pre-trained Weights	Method	L	NL	Acc.
BioCLIP	MSP	90.22	91.61	91.14
	MLS	90.7	92.7	
	EBO	90.58	92.6	
ImageNet-1K	MSP	85.02	85.39	83.59
	MLS	85.59	86.31	
	EBO	85.2	85.98	

Stability. Most post-hoc methods show stable performance across the three training runs, with standard deviations below 5. However, KNN [78], DICE [77], and SHE [101] have much higher standard deviation, as for some runs, the AUROC is smaller than 50, indicating that the OSR scores of positive (open-set) samples are generally lower than negative (closed-set) samples. We present the complete results of the three runs of these methods in the Appendix.

Inference efficiency. For species discovery, OSR methods are expected to be applied to a continuous stream of data over the long term, making inference efficiency a key consideration. Besides, the domain of biodiversity monitoring is often constrained by limited computing resources. Therefore, we also evaluate the efficiency of the post-hoc methods at inference time. OpenMax [6], MDS [49], and RMDS [64] are slower than other methods, and KLM [30] is notably much slower. We list the set-up time and inference time of all post-hoc methods in the Appendix. Setup only needs to be performed once during deployment, while inference must continue to run as new data arrive. The table also indicates which methods require access to training or validation data during setup. This information can be important for end-users, as these data may not always be available, and the requirements to process them can be complicated.

5.2 Training-time Regularization

These methods require further training, but they only utilize the closed-set during training time. For regions with more closed-set data (NE-America and W-Europe), training-time regularization does not improve the performance of local or non-local OSR, but it can slightly improve the performance of non-moth OSR. The opposite holds for regions with a much smaller closed-set (C-America).

Recent post-hoc and training-time regularization methods have been shown [93] to have a substantial improvement over the MSP baseline for coarser-grained OOD detection, such as using ImageNet [67] for ID and iNaturalist [82] for OOD. The inconsistent performance observed with the Open-Insect benchmark highlights the value of evaluating methods with a fine-grained biodiversity dataset focused on semantic shift rather than inferring performance only from more generic OOD datasets.

5.3 Auxiliary Data

Realistic selection of auxiliary data helps. Methods using auxiliary data exhibit an increase in performance for local and non-local OSR in Table 2, especially for C-America, on which there is an increase of 3.28 percentage points (pp). We believe the efficacy of the auxiliary data is due to its realistic nature. For biodiversity monitoring, expert knowledge can be used to choose species that might be similar to open-set species – for example, moths from other regions.

To empirically verify this claim, we train the Energy method for C-America while replacing our auxiliary data with TinyImageNet (TIN) [79], and compare the OSR performance in Table 4. We chose TIN as it is the auxiliary data used by OpenOOD [93] for all 4 open-sets. We find that using TIN reduces OSR performance significantly, by 4.82 pp on local (L) and 5.54 pp on non-local (NL) open-sets. This suggests the importance of realistic auxiliary data selection for training.

Species diversity matters in auxiliary data. We also investigate the impact of species diversity on OSR performance. We fix the total number of images and vary the number of species and images per species. We report results for C-America using the proposed NovelBranch approach with pretrained ImageNet weights (which performs best for the local open-set). We observe that the OSR performance improves as the number of species in the auxiliary dataset increases for local novel species, but has slightly less effect for non-local open-set species (see the Appendix). This suggests that data curation efforts may focus on increasing the number of species.

Table 4: **Ablation study on auxiliary datasets.** We report AUROC for OSR performance and accuracy on the closed-set. AUX is the auxiliary dataset used for training. TIN = TinyImageNet, OI-CA = Open-Insect C-America.

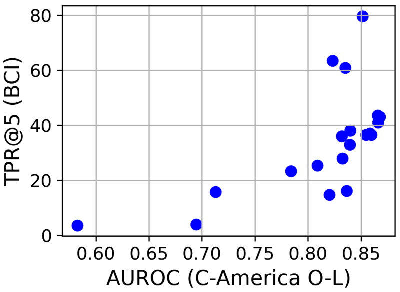
AUX	L	NL	NM	Acc.
TIN	85.25	88.16	93.85	84.17
OI-CA	90.07	93.7	91.33	84.19

5.4 Performance on BCI Data

To test if the performance on the Open-Insect Benchmark can be generalized to likely undescribed species in the wild, we evaluate the post-hoc methods with the BCI images, a highly realistic test case for species discovery. While AUROC captures the overall performance of a model across all thresholds, a specific threshold must be selected to determine whether a species is from the open-set in real-world scenarios.

Given the limited number of taxonomists available to verify new species and the large scale of data collected by camera traps during long-term monitoring, it is crucial to select a threshold that ensures a reasonable False Positive Rate (FPR) while maintaining a high True Positive Rate (TPR), or recall. Here, we consider open-set species as positive and closed-set species as negative. To achieve this goal, we consider the following metrics: the TPR of detecting O-BCI images when the supremum of the FPR of flagging closed-set species as open-set is τ, denoted by TPR@τ. We visualize the result in Fig. 4: the x-axis gives the AUROC of different methods when evaluated on the C-America local open-set, while the y-axis gives the TPR (rate of correctly detecting open-set species), where the threshold is set so that the FPR (misclassifications of closed-set species as open) is lower than 5%. We observe that post-hoc models that perform well on C-America O-L (high AUROC) also tend to achieve higher performance (high TPR@5) on BCI, indicating the OSR performance on Open-Insect is a good approximation for species discovery in the wild.

Figure 4: **TPR@5 (BCI) vs. AUROC (C-America O-L post-hoc methods).** Overall, models that perform well on C-America O-L also tend to achieve higher performance on the BCI data.



5.5 Explainability of the OSR Methods

Explainability of the OSR methods is very important when encouraging domain experts to adopt ML-based tools for species discovery. One common concern is that models only use background features instead of species-level fine-grained features to determine whether the species from the open-set or not. We conducted an experiment to empirically verify that *background features are not enough to achieve good OSR performance on Open-Insect*. Specifically, we masked the insect or the background in a subset of images (see the Appendix for details). We ran our C-America classifier on this subset and evaluated the closed-set classification accuracy and the OSR performance of MSP, a well-performing post-hoc method. We observe that using the original images gives the highest closed-set accuracy (Acc.) and the best OSR performance (AUROC) (Table 5). When the insect was masked, the model was forced to rely solely on the background and the insect's silhouette. Acc. dropped to just 1.63%, and AUROC decreased to 46.57%, close to 50% (the performance of a random classifier). When the background was masked, the model still achieved performance comparable to that on the original images, though Acc. and AUROC decreased by 4% and 6.87%, respectively. When masking the background, parts of the insect such as legs and antennae were also inadvertently masked, which may explain the slight drop.

Table 5: **OSR performance of MPS on different test data types.** Performance dropped drastically when the insect was masked in the image.

Test Data Type	AUROC	Acc.
Original image	80.01	71.01
Moth masked	46.57	1.63
Background masked	76.01	64.14

6 Conclusion

We present the Open-Insect benchmark, a large-scale, fine-grained dataset focused on highly similar insect species. By minimizing covariate shifts of the auxiliary and open-sets from the closed-sets, our dataset allows for a more rigorous evaluation of OSR methods and will provide valuable insights for the future development of these methods. We evaluated 38 OSR methods and found that simple, efficient post-hoc approaches can perform well, achieving over 85% AUROC in detecting fine-grained open-set species. Besides, we show that selecting auxiliary data based on expert knowledge can further improve the OSR performance, with species diversity in the auxiliary dataset playing a key role in performance gains. Our findings also show that methods effective on the Open-Insect benchmark appear to generalize well to images of actual undescribed species.

Positive impacts. We hope that the Open-Insect benchmark will draw attention to the problem of species discovery and enable further work within the ML research community on OSR and OOD detection methods for biodiversity. Such work stands to benefit the biodiversity protection efforts across the world, the fight against climate change (which is exacerbated by biodiversity loss), and the preservation of ecosystem services on which humanity depends.

Potential negative impacts and safeguards. We emphasize that computer vision tools for biodiversity should *not* be seen as a replacement for domain experts in ecology (who are already very limited in number) – rather as a tool to enable such experts to gather and interpret data across more species, geographies, and timepoints than heretofore possible. Besides, classifiers for species recognition and OSR might be misused to identify species of high economic value, which may inadvertently facilitate illegal wildlife trafficking. It is thus important to engage with conservation experts to assess potential misuse before such models are deployed.

Limitations. Our benchmark only evaluates OSR performance on high-resolution images and does not include lower-resolution images, such as those from camera traps, due to the limited availability of annotated trap data. Furthermore, due to limited computational resources, we do not perform multiple training runs for "training regularization" and "training with auxiliary data" methods.

Acknowledgments

This work was supported by the AI and Biodiversity Change (ABC) Global Center, which is funded by the US National Science Foundation (NSF) under Award No. 2330423 and Natural Sciences and Engineering Research Council (NSERC) of Canada under Award No. 585136. The work draws on research supported by the Social Sciences and Humanities Research Council (SSHRC). Any opinions, findings and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the NSF, NSERC, or SSHRC. This work was also supported in part by the Pioneer Centre for AI, DNRF grant number P1, by the Global Wetland Center (grant number NNF23OC0081089) from Novo Nordisk Foundation, by the Canada CIFAR AI Chairs program, by the Fonds de recherche du Québec – Nature et technologies (FRQNT), by the MIT-IBM Watson AI Lab Partners Program, by the MIT-Google Program for Computing Innovation, by the MIT MGAIC Seed Funds program, by NSF Award No. 2441060, and by the Schmidt Sciences AI2050 Program.

This research was enabled in part by compute resources, software and technical help provided by Mila (mila.quebec). We thank Francis Pelletier for the valuable guidance on the code implementation and dataset preparation.

References

[1] GBIF.org (24 July 2024) GBIF Occurrence Download. `https://doi.org/10.15468/dl.jzy3de`. 26

[2] GBIF.org (24 July 2024) GBIF Occurrence Download . `https://doi.org/10.15468/dl.6j5bzj`. 26

[3] Mouïn Ben Ammar, Nacim Belkhir, Sebastian Popescu, Antoine Manzanera, and Gianni Franchi. NECO: NEural collapse based out-of-distribution detection. In *The Twelfth International Conference on Learning Representations*, 2024. 3, 6, 7, 34

[4] Haoyue Bai, Gregory Canal, Xuefeng Du, Jeongyeol Kwon, Robert D Nowak, and Yixuan Li. Feed two birds with one scone: Exploiting wild data for both out-of-distribution generalization and detection. In *International Conference on Machine Learning*, pages 1454–1471. PMLR, 2023. 3

[5] Haoyue Bai, Jifan Zhang, and Robert D Nowak. AHA: Human-assisted out-of-distribution generalization and detection. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024. 3

[6] Abhijit Bendale and Terrance E Boult. Towards open set deep networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1563–1572, 2016. 3, 7, 8, 34

[7] Julian Bitterwolf, Maximilian Müller, and Matthias Hein. In or out? fixing imagenet out-of-distribution detection evaluation. In *Proceedings of the 40th International Conference on Machine Learning*, pages 2471–2506, 2023. 2

[8] Kim Bjerge, Jakob Bonde Nielsen, Martin Videbæk Sepstrup, Flemming Helsing-Nielsen, and Toke Thomas Høye. An automated light trap to monitor moths (lepidoptera) using computer vision-based tracking and deep learning. *Sensors*, 21(2):343, 2021. 4

[9] O. Bánki, Y. Roskov, M. Döring, G. Ower, D. R. Hernández Robles, C. A. Plata Corredor, T. Stjernegaard Jeppesen, A. Örn, T. Pape, D. Hobern, S. Garnett, H. Little, R. E. DeWalt, K. Ma, J. Miller, T. Orrell, R. Aalbu, J. Abbott, C. Aedo, and et al. Catalogue of life (version 2025-01-17), 2025. 2

[10] Vojtěch Čermák, Lukas Picek, Lukáš Adam, and Kostas Papafitsoros. Wildlifedatasets: An open-source toolkit for animal re-identification. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 5953–5963, 2024. 2

[11] 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):8065–8081, 2021. 3, 6, 7

[12] Steve Cruz, Ryan Rabinowitz, Manuel Günther, and Terrance E Boult. Operational open-set recognition and postmax refinement. In *European Conference on Computer Vision*, pages 475–492. Springer, 2024. 7, 28

[13] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009. 2, 3

[14] Li Deng. The mnist database of handwritten digit images for machine learning research [best of the web]. *IEEE signal processing magazine*, 29(6):141–142, 2012. 3

[15] Terrance DeVries and Graham W Taylor. Learning confidence for out-of-distribution detection in neural networks. *arXiv preprint arXiv:1802.04865*, 2018. 3, 6, 7

[16] Akshay Raj Dhamija, Manuel Günther, and Terrance Boult. Reducing network agnostophobia. *Advances in Neural Information Processing Systems*, 31, 2018. 3, 7, 28

[17] Andrija Djurisic, Nebojsa Bozanic, Arjun Ashok, and Rosanne Liu. Extremely simple activation shaping for out-of-distribution detection. In *The Eleventh International Conference on Learning Representations*, 2023. 3, 7, 34

[18] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations*, 2021. 7, 36

[19] Xuefeng Du, Zhen Fang, Ilias Diakonikolas, and Yixuan Li. How does unlabeled data provably help out-of-distribution detection? In *The Twelfth International Conference on Learning Representations*, 2024. 3

[20] Xuefeng Du, Zhaoning Wang, Mu Cai, and Yixuan Li. Vos: Learning what you don't know by virtual outlier synthesis. *Proceedings of the International Conference on Learning Representations*, 2022. 3

[21] Ke Fan, Tong Liu, Xingyu Qiu, Yikai Wang, Lian Huai, Zeyu Shangguan, Shuang Gou, Fengjian Liu, Yuqian Fu, Yanwei Fu, et al. Test-time linear out-of-distribution detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 23752–23761, 2024. 3

[22] Ido Galil, Mohammed Dabbah, and Ran El-Yaniv. A framework for benchmarking class-out-of-distribution detection and its application to imagenet. In *The Eleventh International Conference on Learning Representations*, 2023. 2, 3

[23] Zahra Gharaee, ZeMing Gong, Nicholas Pellegrino, Iuliia Zarubiieva, Joakim Bruslund Haurum, Scott Lowe, Jaclyn McKeown, Chris Ho, Joschka McLeod, Yi-Yun Wei, Jireh Agda, Sujeevan Ratnasingham, Dirk Steinke, Angel Chang, Graham W Taylor, and Paul Fieguth. A Step Towards Worldwide Biodiversity Assessment: The BIOSCAN-1M Insect Dataset. In A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine, editors, *Advances in Neural Information Processing Systems*, volume 36, pages 43593–43619. Curran Associates, Inc., 2023. 2, 4, 26

[24] Zahra Gharaee, Scott C. Lowe, ZeMing Gong, Pablo Millan Arias, Nicholas Pellegrino, Austin T. Wang, Joakim Bruslund Haurum, Iuliia Zarubiieva, Lila Kari, Dirk Steinke, Graham W. Taylor, Paul Fieguth, and Angel X. Chang. BIOSCAN-5M: A multimodal dataset for insect biodiversity. In A. Globerson, L. Mackey, D. Belgrave, A. Fan, U. Paquet, J. Tomczak, and C. Zhang, editors, *Advances in Neural Information Processing Systems*, volume 37, pages 36285–36313. Curran Associates, Inc., 2024. 2, 4, 26

[25] ZeMing Gong, Austin Wang, Xiaoliang Huo, Joakim Bruslund Haurum, Scott C. Lowe, Graham W. Taylor, and Angel X Chang. CLIBD: Bridging vision and genomics for biodiversity monitoring at scale. In *The Thirteenth International Conference on Learning Representations*, 2025. 4

[26] Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q Weinberger. On calibration of modern neural networks. In *International conference on machine learning*, pages 1321–1330. PMLR, 2017. 3, 7, 8, 34

[27] 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. 6

[28] Tong He, Zhi Zhang, Hang Zhang, Zhongyue Zhang, Junyuan Xie, and Mu Li. Bag of tricks for image classification with convolutional neural networks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 558–567, 2019. 6, 28

[29] 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*, pages 8759–8773. PMLR, 2022. 3, 7, 34

[30] 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 *Proceedings of the 39th International Conference on Machine Learning*, volume 162, pages 8759–8773, 2022. 3, 7, 8, 34

[31] 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. 3, 7, 8, 34

[32] Dan Hendrycks, Mantas Mazeika, and Thomas Dietterich. Deep anomaly detection with outlier exposure. In *International Conference on Learning Representations*, 2019. 3, 7

[33] Dan Hendrycks, Mantas Mazeika, Saurav Kadavath, and Dawn Song. Using self-supervised learning can improve model robustness and uncertainty. *Advances in neural information processing systems*, 32, 2019. 3, 6, 7

[34] Dan Hendrycks, Kevin Zhao, Steven Basart, Jacob Steinhardt, and Dawn Song. Natural adversarial examples. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 15262–15271, 2021. 2, 3

[35] Helmut Hillebrand. On the generality of the latitudinal diversity gradient. *The American Naturalist*, 163(2):192–211, 2004. 26

[36] Claus Hofmann, Simon Lucas Schmid, Bernhard Lehner, Daniel Klotz, and Sepp Hochreiter. Energy-based hopfield boosting for out-of-distribution detection. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024. 3

[37] Laurens E Hogeweg, Rajesh Gangireddy, Django Brunink, Vincent J Kalkman, Ludo Cornelissen, and Jacob W Kamminga. Cood: Combined out-of-distribution detection using multiple measures for anomaly & novel class detection in large-scale hierarchical classification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3971–3980, 2024. 2, 3, 5

[38] 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. 3, 6, 7

[39] Rui Huang, Andrew Geng, and Yixuan Li. On the importance of gradients for detecting distributional shifts in the wild. *Advances in Neural Information Processing Systems*, 34:677–689, 2021. 3, 7, 34

[40] Rui Huang and Yixuan Li. Mos: Towards scaling out-of-distribution detection for large semantic space. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8710–8719, 2021. 3

[41] Aditya Jain, Fagner Cunha, Michael James Bunsen, Juan Sebastián Cañas, Léonard Pasi, Nathan Pinoy, Flemming Helsing, JoAnne Russo, Marc Botham, Michael Sabourin, Jonathan Fréchette, Alexandre Anctil, Yacksecari Lopez, Eduardo Navarro, Filonila Perez Pimentel, Ana Cecilia Zamora, José Alejandro Ramirez Silva, Jonathan Gagnon, Tom August, Kim Bjerge, Alba Gomez Segura, Marc Bélisle, Yves Basset, Kent P. McFarland, David Roy, Toke Thomas Høye, Maxim Larrivée, and David Rolnick. Insect identification innbsp;thenbsp;wild: The ami dataset. In *Computer Vision – ECCV 2024: 18th European Conference, Milan, Italy, September 29–October 4, 2024, Proceedings, Part XXXVII*, page 55–73, Berlin, Heidelberg, 2024. Springer-Verlag. 4, 23, 26

[42] Wenyu Jiang, Hao Cheng, MingCai Chen, Chongjun Wang, and Hongxin Wei. DOS: Diverse outlier sampling for out-of-distribution detection. In *The Twelfth International Conference on Learning Representations*, 2024. 3

[43] Xue Jiang, Feng Liu, Zhen Fang, Hong Chen, Tongliang Liu, Feng Zheng, and Bo Han. Detecting out-of-distribution data through in-distribution class prior. In *International Conference on Machine Learning*, pages 15067–15088. PMLR, 2023. 3, 6, 7, 8, 34

[44] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C Berg, Wan-Yen Lo, et al. Segment anything. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 4015–4026, 2023. 35

[45] Shu Kong and Deva Ramanan. Opengan: Open-set recognition via open data generation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 813–822, 2021. 3, 6, 7

[46] Alex Krizhevsky, Geoffrey Hinton, et al. Learning multiple layers of features from tiny images. 2009. 3

[47] Gustaf Kylberg. The kylberg texture dataset v. 1.0. Technical Report External report (Blue series) No. 35, Centre for Image Analysis, Swedish University of Agricultural Sciences and Uppsala University, 2011. 3

[48] Nico Lang, Vésteinn Snæbjarnarson, Elijah Cole, Oisin Mac Aodha, Christian Igel, and Serge Belongie. From coarse to fine-grained open-set recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 17804–17814, 2024. 3, 5, 6, 8

[49] Kimin Lee, Kibok Lee, Honglak Lee, and Jinwoo Shin. A simple unified framework for detecting out-of-distribution samples and adversarial attacks. In *Proceedings of the 32nd International Conference on Neural Information Processing Systems*, pages 7167–7177, 2018. 3, 7, 8, 34

[50] 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. 3, 7, 34

[51] Litian Liu and Yao Qin. Fast decision boundary based out-of-distribution detector. In *Forty-first International Conference on Machine Learning*, 2024. 3, 6, 7, 34

[52] Litian Liu and Yao Qin. Detecting out-of-distribution through the lens of neural collapse. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 15424–15433, 2025. 3, 6, 7, 34

[53] Shilong Liu, Zhaoyang Zeng, Tianhe Ren, Feng Li, Hao Zhang, Jie Yang, Qing Jiang, Chunyuan Li, Jianwei Yang, Hang Su, et al. Grounding dino: Marrying dino with grounded pre-training for open-set object detection. In *European conference on computer vision*, pages 38–55. Springer, 2024. 35

[54] Weitang Liu, Xiaoyun Wang, John D Owens, and Yixuan Li. Energy-based out-of-distribution detection. In *Proceedings of the 34th International Conference on Neural Information Processing Systems*, pages 21464–21475, 2020. 3, 6, 7, 34

[55] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *International Conference on Learning Representations*, 2019. 6, 28

[56] Alexandra Sasha Luccioni and David Rolnick. Bugs in the data: How imagenet misrepresents biodiversity. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pages 14382–14390, 2023. 2

[57] Wenjun Miao, Guansong Pang, Jin Zheng, and Xiao Bai. Long-tailed out-of-distribution detection via normalized outlier distribution adaptation. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024. 3

[58] Yifei Ming, Yiyou Sun, Ousmane Dia, and Yixuan Li. How to exploit hyperspherical embeddings for out-of-distribution detection? In *The Eleventh International Conference on Learning Representations*, 2023. 3

[59] Camilo Mora, Derek P Tittensor, Sina Adl, Alastair GB Simpson, and Boris Worm. How many species are there on Earth and in the ocean? *PLoS biology*, 9(8):e1001127, 2011. 2

[60] Harikrishna Narasimhan, Aditya Krishna Menon, Wittawat Jitkrittum, and Sanjiv Kumar. Plugin estimators for selective classification with out-of-distribution detection. In *The Twelfth International Conference on Learning Representations*, 2024. 3

[61] Hoang-Quan Nguyen, Thanh-Dat Truong, Xuan Bac Nguyen, Ashley Dowling, Xin Li, and Khoa Luu. Insect-foundation: A foundation model and large-scale 1m dataset for visual insect understanding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 21945–21955, 2024. 4, 26

[62] Sujeevan Ratnasingham and Paul DN Hebert. Bold: The barcode of life data system (http://www. barcodinglife. org). *Molecular ecology notes*, 7(3):355–364, 2007. 5

[63] Sujeevan Ratnasingham and Paul DN Hebert. A dna-based registry for all animal species: the barcode index number (bin) system. *PloS one*, 8(7):e66213, 2013. 5

[64] Jie Ren, Stanislav Fort, Jeremiah Liu, Abhijit Guha Roy, Shreyas Padhy, and Balaji Lakshminarayanan. A simple fix to mahalanobis distance for improving near-ood detection. *arXiv preprint arXiv:2106.09022*, 2021. 3, 7, 8, 34

[65] László Ronkay and Peter Huemer. Agrotis fatidica (hübner, 1824) species-group revisited, with description of two new species from the alps and the pyrenees (lepidoptera, noctuidae). *Nota Lepidopterologica*, 41:145–179, 2018. 4

[66] DB Roy, J Alison, TA August, M Bélisle, K Bjerge, JJ Bowden, MJ Bunsen, F Cunha, Q Geissmann, K Goldmann, et al. Towards a standardized framework for ai-assisted, image-based monitoring of nocturnal insects. *Philosophical Transactions of the Royal Society B*, 379(1904):20230108, 2024. 4

[67] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, et al. Imagenet large scale visual recognition challenge. *International journal of computer vision*, 115:211–252, 2015. 2, 3, 8

[68] Chandramouli Shama Sastry and Sageev Oore. Detecting out-of-distribution examples with Gram matrices. In Hal Daumé III and Aarti Singh, editors, *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pages 8491–8501. PMLR, 13–18 Jul 2020. 3, 7, 34

[69] 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):1757–1772, 2012. 3

[70] Maarten Schermer and Laurens Hogeweg. Supporting citizen scientists with automatic species identification using deep learning image recognition models. *Biodiversity Information Science and Standards*, 2018. 3

[71] Sina Sharifi, Taha Entesari, Bardia Safaei, Vishal M Patel, and Mahyar Fazlyab. Gradient-regularized out-of-distribution detection. In *European Conference on Computer Vision*, pages 459–478. Springer, 2025. 3

[72] Sanjay Sondhi, Tarun Karmakar, Yash Sondhi, and Krushnamegh Kunte. Moths of tale wildlife sanctuary, arunachal pradesh, india with seventeen additions to the moth fauna of india (lepidoptera: Heterocera). *Tropical Lepidoptera Research*, pages 1–53, 2021. 4

[73] Yue Song, Nicu Sebe, and Wei Wang. Rankfeat: rank-1 feature removal for out-of-distribution detection. In *Proceedings of the 36th International Conference on Neural Information Processing Systems*, pages 17885–17898, 2022. 3, 7, 34

[74] Samuel Stevens, Jiaman Wu, Matthew J Thompson, Elizabeth G Campolongo, Chan Hee Song, David Edward Carlyn, Li Dong, Wasila M Dahdul, Charles Stewart, Tanya Berger-Wolf, et al. Bioclip: A vision foundation model for the tree of life. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 19412–19424, 2024. 2, 3, 4, 7, 35, 36

[75] Nigel E Stork. How many species of insects and other terrestrial arthropods are there on earth? *Annual review of entomology*, 63(1):31–45, 2018. 2

[76] Yiyou Sun, Chuan Guo, and Yixuan Li. React: out-of-distribution detection with rectified activations. In *Proceedings of the 35th International Conference on Neural Information Processing Systems*, pages 144–157, 2021. 3, 7, 34

[77] Yiyou Sun and Yixuan Li. Dice: Leveraging sparsification for out-of-distribution detection. In *European Conference on Computer Vision*, pages 691–708. Springer, 2022. 3, 7, 8, 34

[78] Yiyou Sun, Yifei Ming, Xiaojin Zhu, and Yixuan Li. Out-of-distribution detection with deep nearest neighbors. In *International Conference on Machine Learning*, pages 20827–20840. PMLR, 2022. 3, 7, 8, 34

[79] Antonio Torralba, Rob Fergus, and William T Freeman. 80 million tiny images: A large data set for nonparametric object and scene recognition. *IEEE transactions on pattern analysis and machine intelligence*, 30(11):1958–1970, 2008. 3, 9

[80] Devis Tuia, Benjamin Kellenberger, Sara Beery, Blair R Costelloe, Silvia Zuffi, Benjamin Risse, Alexander Mathis, Mackenzie W Mathis, Frank Van Langevelde, Tilo Burghardt, et al. Perspectives in machine learning for wildlife conservation. *Nature communications*, 13(1):1–15, 2022. 2

[81] Grant Van Horn, Elijah Cole, Sara Beery, Kimberly Wilber, Serge Belongie, and Oisin Mac Aodha. Benchmarking representation learning for natural world image collections. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 12884–12893, 2021. 2, 3, 6, 36

[82] Grant Van Horn, Oisin Mac Aodha, Yang Song, Yin Cui, Chen Sun, Alex Shepard, Hartwig Adam, Pietro Perona, and Serge Belongie. The inaturalist species classification and detection dataset. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 8769–8778, 2018. 2, 3, 8

[83] Roel Van Klink, Tom August, Yves Bas, Paul Bodesheim, Aletta Bonn, Frode Fossøy, Toke T Høye, Eelke Jongejans, Myles HM Menz, Andreia Miraldo, et al. Emerging technologies revolutionise insect ecology and monitoring. *Trends in ecology & evolution*, 37(10):872–885, 2022. 4

[84] 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*, 2022. 2, 3, 5, 7, 8, 28, 36

[85] Edward Vendrow, Omiros Pantazis, Alexander Shepard, Gabriel Brostow, Kate E. Jones, Oisin Mac Aodha, Sara Beery, and Grant Van Horn. INQUIRE: A natural world text-to-image retrieval benchmark. In *The Thirty-eight Conference on Neural Information Processing Systems Datasets and Benchmarks Track*, 2024. 2

[86] Catherine Wah, Steve Branson, Peter Welinder, Pietro Perona, and Serge Belongie. The caltech-ucsd birds-200-2011 dataset. Technical report, August 2023. 3

[87] Haoqi Wang, Zhizhong Li, Litong Feng, and Wayne Zhang. Vim: Out-of-distribution with virtual-logit matching. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 4921–4930, 2022. 3, 7, 34

[88] Zitai Wang, Qianqian Xu, Zhiyong Yang, Yuan He, Xiaochun Cao, and Qingming Huang. Openauc: Towards auc-oriented open-set recognition. *Advances in Neural Information Processing Systems*, 35:25033–25045, 2022. 7, 28

[89] Hongxin Wei, Renchunzi Xie, Hao Cheng, Lei Feng, Bo An, and Yixuan Li. Mitigating neural network overconfidence with logit normalization. In *International conference on machine learning*, pages 23631–23644. PMLR, 2022. 3, 6, 7

[90] Chenhui Xu, Fuxun Yu, Zirui Xu, Nathan Inkawhich, and Xiang Chen. Out-of-distribution detection via deep multi-comprehension ensemble. In *Proceedings of the 41st International Conference on Machine Learning*, pages 55465–55489, 2024. 3

[91] Chih-Hsuan Yang, Benjamin Feuer, Talukder Zaki Jubery, Zi K. Deng, Andre Nakkab, Md Zahid Hasan, Shivani Chiranjeevi, Kelly O. Marshall, Nirmal Baishnab, Asheesh K Singh, ARTI SINGH, Soumik Sarkar, Nirav Merchant, Chinmay Hegde, and Baskar Ganapathysubramanian. BioTrove: A Large Curated Image Dataset Enabling AI for Biodiversity. In *The Thirty-eight Conference on Neural Information Processing Systems Datasets and Benchmarks Track*, 2024. 2, 3, 4

[92] Jingkang Yang, Haoqi Wang, Litong Feng, Xiaopeng Yan, Huabin Zheng, Wayne Zhang, and Ziwei Liu. Semantically coherent out-of-distribution detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 8301–8309, 2021. 7

[93] Jingkang Yang, Pengyun Wang, Dejian Zou, Zitang Zhou, Kunyuan Ding, Wenxuan Peng, Haoqi Wang, Guangyao Chen, Bo Li, Yiyou Sun, et al. Openood: Benchmarking generalized out-of-distribution detection. *Advances in Neural Information Processing Systems*, 35:32598–32611, 2022. 3, 6, 8, 9, 23, 28

[94] Jingkang Yang, Kaiyang Zhou, Yixuan Li, and Ziwei Liu. Generalized out-of-distribution detection: A survey. *International Journal of Computer Vision*, 132(12):5635–5662, 2024. 3

[95] William Yang, Byron Zhang, and Olga Russakovsky. Imagenet-OOD: Deciphering modern out-of-distribution detection algorithms. In *The Twelfth International Conference on Learning Representations*, 2024. 2, 3

[96] Haiyun Yao, Zongbo Han, Huazhu Fu, Xi Peng, Qinghua Hu, and Changqing Zhang. Out-of-distribution detection with diversification (provably). In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024. 3

[97] Qing Yu and Kiyoharu Aizawa. Unsupervised out-of-distribution detection by maximum classifier discrepancy. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 9518–9526, 2019. 6

[98] Reza Zahiri, J Donald Lafontaine, B Christian Schmidt, Jeremy R Dewaard, Evgeny V Zakharov, and Paul DN Hebert. Probing planetary biodiversity with dna barcodes: The noctuoidea of north america. *PloS one*, 12(6):e0178548, 2017. 5

[99] Jingyang Zhang, Nathan Inkawhich, Randolph Linderman, Yiran Chen, and Hai Li. Mixture outlier exposure: Towards out-of-distribution detection in fine-grained environments. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 5531–5540, 2023. 7

[100] Jingyang Zhang, Jingkang Yang, Pengyun Wang, Haoqi Wang, Yueqian Lin, Haoran Zhang, Yiyou Sun, Xuefeng Du, Yixuan Li, Ziwei Liu, Yiran Chen, and Hai Li. OpenOOD v1.5: Enhanced benchmark for out-of-distribution detection. *Journal of Data-centric Machine Learning Research*, 2024. Dataset Certification. 3, 6, 7

[101] Jinsong Zhang, Qiang Fu, Xu Chen, Lun Du, Zelin Li, Gang Wang, Shi Han, Dongmei Zhang, et al. Out-of-distribution detection based on in-distribution data patterns memorization with modern hopfield energy. In *The Eleventh International Conference on Learning Representations*, 2022. 3, 7, 8, 34

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The main claims made in the abstract and introduction summarize the paper's contributions and scope.

   Guidelines:

   - The answer NA means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: We discussion the limitations in Sec. 6.

   Guidelines:

   - The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [NA]

Justification: This paper does not include theorems.

Guidelines:

- The answer NA means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

Answer: [Yes]

Justification: We provide training details in Sec 4 as well as App. B.1. Our code is available at `https://github.com/RolnickLab/Open-Insect`. Our datasets are avaiable at `https://huggingface.co/datasets/yuyan-chen/open-insect` and `https://huggingface.co/datasets/yuyan-chen/open-insect-bci`.

Guidelines:

- The answer NA means that the paper does not include experiments.
- If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
- If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
- Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
- While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
  (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
  (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
  (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
  (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

Answer: [Yes]

Justification: Our datasets are hosted on huggingface: `https://huggingface.co/datasets/yuyan-chen/open-insect` and `https://huggingface.co/datasets/yuyan-chen/open-insect-bci`. Our code is available at `https://github.com/RolnickLab/Open-Insect`.

Guidelines:

- The answer NA means that paper does not include experiments requiring code.
- Please see the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- While we encourage the release of code and data, we understand that this might not be possible, so "No" is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

Answer: [Yes]

Justification: We provide training details in Sec 4 as well as App. B.1. Our code is available at `https://github.com/RolnickLab/Open-Insect`.

Guidelines:

- The answer NA means that the paper does not include experiments.
- The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
- The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

Answer: [Yes]

Justification: The main results are reported in Table 2. We report the mean and standard deviation for post-hoc methods from three training runs. Other training methods were only trained once, and we discuss this limitation in Sec. 6.

Guidelines:

- The answer NA means that the paper does not include experiments.
- The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.

- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
- If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes] ,

   Justification: We provide these information in Sec. B.1, C.3, and 4.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: The research conducted in the paper conforms with the NeurIPS Code of Ethics.

   Guidelines:

   - The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [Yes]

    Justification: We discuss both positive and negative impacts in Sec. 6.

    Guidelines:

    - The answer NA means that there is no societal impact of the work performed.

- If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.
- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

    Answer: [Yes]

    Justification: We discuss safeguards in Sec. 6.

    Guidelines:

    - The answer NA means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: Our codebase is built using OpenOOD [93]. We cite the paper of OpenOOD and include the URL of the GitHub repository in Sec. 4. Our dataset partially builds upon data from the AMI dataset [41]. We cite this paper in Sec. 3.1. We include the license of the AMI dataset in the Open-Insect dataset hosting on huggingface `https://huggingface.co/datasets/yuyan-chen/open-insect`.

    Guidelines:

    - The answer NA means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.

- The name of the license (e.g., CC-BY 4.0) should be included for each asset.
- For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [Yes]

    Justification: Our datasets are hosted on huggingface: `https://huggingface.co/datasets/yuyan-chen/open-insect` and `https://huggingface.co/datasets/yuyan-chen/open-insect-bci`. Our code is available at `https://github.com/RolnickLab/Open-Insect`. Details about training and limitations are discussed in the main text and the Appendix. Both datasets are distributed under the CC-BY-NC 4.0 License. The code is distributed under the MIT License.

    Guidelines:

    - The answer NA means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [NA]

    Justification: This paper does not involve crowdsourcing nor research with human subjects.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [NA]

Justification: This paper does not involve crowdsourcing nor research with human subjects.

Guidelines:

- The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
- Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does not impact the core methodology, scientific rigorousness, or originality of the research, declaration is not required.

Answer: [NA]

Justification: This research does not involve LLMs.

Guidelines:

- The answer NA means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
- Please refer to our LLM policy (`https://neurips.cc/Conferences/2025/LLM`) for what should or should not be described.

Technical Appendices and Supplementary Material

A Dataset

A.1 Biome list

1: “Tropical and Subtropical Moist Broadleaf Forests", 2: “Tropical and Subtropical Dry Broadleaf Forests", 3: “Tropical and Subtropical Coniferous Forests", 4: “Temperate Broadleaf and Mixed Forests", 5: “Temperate Coniferous Forests", 6: “Boreal Forests/Taiga", 7: “Tropical and Subtropical Grasslands, Savannas, and Shrublands“, 8: “Temperate Grasslands, Savannas, and Shrublands", 9: “Flooded Grasslands and Savannas", 10: “Montane Grasslands and Shrublands", 11: “Tundra", 12: “Mediterranean Forests, Woodlands, and Scrub", 13: “Deserts and Xeric Shrublands", 14: “Mangroves", 98: “Lake", 99: “Rock and Ice".

A.2 Choice of degrees in latitude and longitude

We wanted to choose surrounding areas that are small enough so that the species can simulate undescribed species, but at the same time, large enough so that we have a relatively diverse species distribution. Since typically the variation in species is higher across latitudes than longitudes due to climatic shifts [35], we chose 1 degree in latitude and 3 degrees in longitude. We also visualize the biomes of each region in Fig. 2 to show that the surrounding regions are very similar to the in-distribution regions.

A.3 Global moth data curation

We curate the global list of moth species from GBIF [1]. Together, the moth and butterfly families make up the order *Lepidoptera*. While there are over 100 families of moths, there are only seven families of butterflies. Therefore, we exclude the butterfly families from the order Lepidoptera to get only moth families and their corresponding species. Next, we consider only those species whose taxonomic status is classified as accepted in the backbone. This process results in a total of 46,983 species. Simultaneously, we download metadata for all Lepidoptera observations that include images using GBIF's occurrence search tool [2]. We then fetch images for the moth species, limiting to 1,000 images per species class. This effort yields 4.5 million images. Following the data cleaning procedures outlined in [41], we remove images with duplicate URLs, problematic sources, thumbnails, and non-adult images using a life stage classifier. After these steps, we have 3.8 million images remaining. Additionally, we eliminate any species that have fewer than five images in the dataset and only include images with the following licenses: CC BY 4.0, CC BY-NC 4.0, CC BY-NC-SA 4.0, CC0 1.0 (Public Domain), CC BY-SA 4.0, No Rights Reserved, CC BY 3.0, CC BY-SA 3.0. This finalizes our globla data to have 28,388 species and 2,912,168 images.

A.4 The Open-Insect Dataset

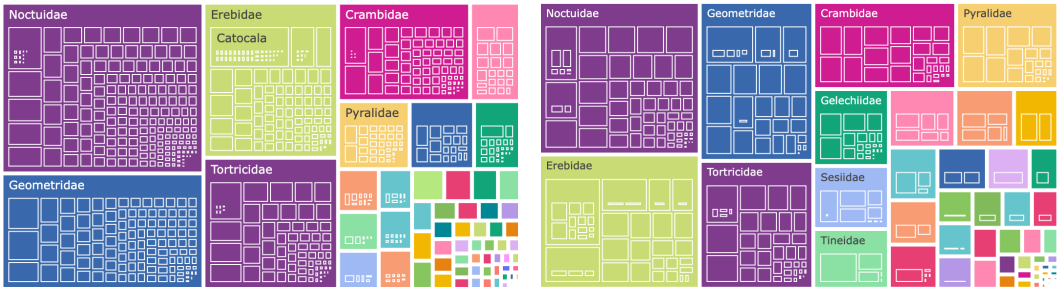
We compare the data distribution of local and non-local open-set datasets of three regions: Northeastern North America (NE-America), Western Europe (W-Europe), and Central America (C-America) in Table 6 and visualize the taxonomic distribution in Fig. 5.

Table 6: **Geographic and taxonomic distance of Open-Insect open-set splits.** Comparison of the local and non-local open-set datasets.

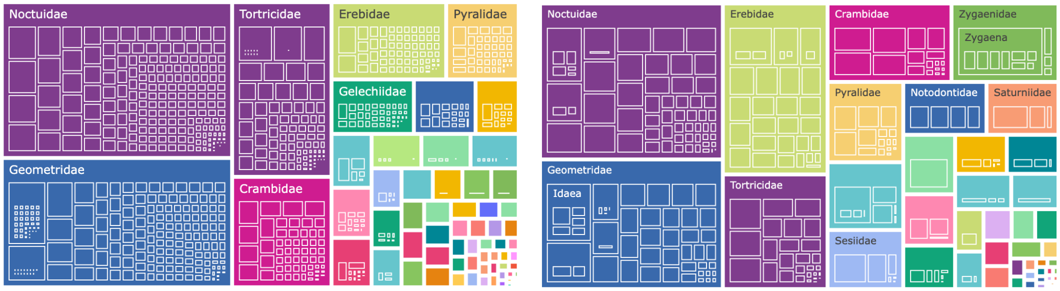
	NE-America				W-Europe				C-America			
	Local		Non-local		Local		Non-local		Local		Non-local	
	Species	Images	Species	Images	Species	Images	Species	Images	Species	Images	Species	Images
1-hop	463	66,370	553	11,919	323	44,119	802	16,962	772	46,731	151	3,428
2-hop	147	46,407	3,810	77,967	132	33,796	3,482	71,355	908	61,651	3,238	68,508
3-hop	7	857	299	5,711	3	733	378	7,280	44	3,497	1,273	23,661

A.5 Comparison to related work

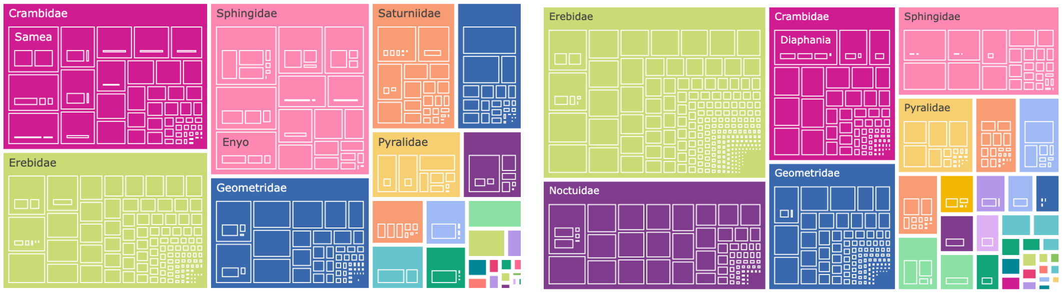
Taxonomic level. BIOSCAN-1M (B-1M) [23], BIOSCAN-5M (B-5M) [24], and our Open-Insect, focus on insects, while Insect-1M [61] consist of insects and other arthropods. B-1M and B-5M



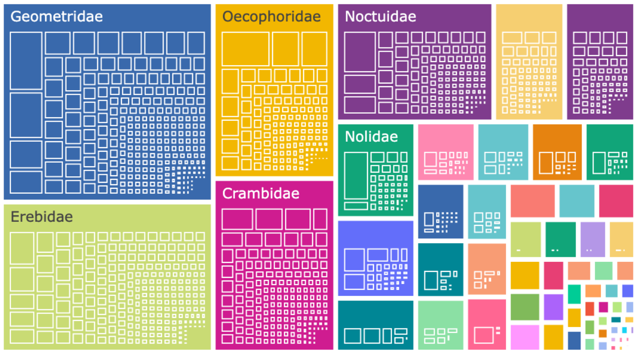
(a) NE-America Closed-Set (b) NE-America Open-Set Local



(c) W-Europe Closed-Set (d) W-Europe Open-Set Local



(e) C-America Closed-Set (f) C-America Open-Set Local



(g) Non-Local Open-Set

Figure 5: **Visualization of the Open-Insect taxonomic distribution.** Tree maps (a)–(f) show the taxonomic composition of moth families in Open-Insect across three regions. Each nested box represents a genus or species, and box size reflects the relative number of images. The same family is *colored consistently* across regions. Local open-set species display taxonomic distributions more similar to their corresponding closed-set species, indicating shared families and comparable visual traits. In contrast, the non-local open-set samples from Australia (g) exhibit markedly different taxonomic and color patterns, reflecting greater divergence from the training regions.

mainly focus on flies and use Barcode Index Numbers (BINs) as labels (< 10% of these BINs correspond to species, while others are higher order taxonomic groups of various levels). Our dataset focuses on moths and all samples are identified to species level. This allows us to evaluate the difficulty of OSR in highly fine-grained recognition and at specified taxonomic levels.

Data type. While all datasets include images, the image types are different. Open-Insect and Insect-1M contain images of museum specimens and live insects in the wild. B-1M and B-5M consist of microscope images of specimens collected using Malaise traps. While Malaise trap specimens are often too small and mangled for species-level identification, Open-Insect was designed to support species-level identification and new species discovery, focusing on more readily identifiable camera trap images, as well as slightly larger insects that can be visually identified.

Benchmark tasks: The intended applications of the other three datasets are very different from those of Open-Insect. B-1M and B-5M assess the effectiveness of multi-modality by aligning image data and DNA barcodes, while Insect-1M was curated to train a foundation model for insects and other arthropods using images and text. Open-Insect is designed for visual species identification where DNA barcodes or text description may not be available. Given the strong covariate shift and difference in tasks between B-1M / B-5M and Open-Insect, it is not possible to directly evaluate OSR performance of models trained on B-1M and B-5M using Open-Insect.

B Methods

B.1 Training details

All images were resized to 126 by 126. We followed the data pre-processing implementation in OpenOOD [93]. All models were trained with 1 RTX8000 GPU, with 16 CPUs, 16 workers, and 100 GB of CPU memory in total.

Training from scratch. All species recognition classifiers were trained from scratch for 120 epochs with a batch size of 512 for each region. We used the AdamW optimizer [55] with an initial learning rate of 0.01. The learning rate was decayed with cosine scheduling [28] with the first 6 epochs being warm-up epochs.

Fine-tuning. For all training methods (whether requiring extra data or not) , we fine-tuned the species-level classifiers for 30 additional epochs with a batch size of 512 except for RotPred, which was trained with a batch size of 256 due to a larger memory requirement. Similarly, we used AdamW as the optimizer with cosine scheduling for learning rate scheduling. The initial learning rate was 0.001, with the first two epochs as warm-up.

C Additional results

C.1 All metrics

We show the Open-Set Classification Rate (OSCR0 curve [16] of the basic classifier with MSP as the post-hoc method across the three regions in Fig. 6. We also report the Area Under the OSCR curve (AUOSCR) [84], OpenAUC [88], and Operational Open-Set Accuracy (OOSA) [12] in Tables 7, 8, and 9. AUOSCR, OpenAUC, and OOSA depend on the closed-set accuracy, whereas AUROC is independent of it. The closed-set accuracies reported in Table 2 are computed in the standard manner, using the argmax of the logit vector. Hence, all post-hoc methods have the same closed-set classification accuracy. However, some post-hoc methods also provide an alternative way to derive predictions. In such cases, we report the metrics computed using both (1) the standard approach and (2) the post-hoc prediction method (denoted with *) in the subsequent tables if the metrics are drastically different.

C.2 OSR performance when using ImageNet-1K pretrained weights

We present the result in Table 10. There is no result for OpenMax as it requires prediction of the test set to cover all training species. Though the ground truths of the test set cover all training species, the predictions of this model do not.

Table 7: **Full results of NE-America.** AUROC, AUOSCR, OpenAUC, and OOSA are averaged across three open-set splits: Local, Non-Local, and Non-Moth. For post-hoc methods, we present the result of one of the three runs.

Method	AUROC	AUOSCR	OpenAUC	OOSA	ACC
MSP	91.7	85.1	85.1	84.5	89.8
TempScale	91.7	85.1	85.1	84.5	89.8
ODIN	88.9	81.9	81.9	81.0	89.7
MDS	82.2	77.5	77.5	77.8	89.8
MDSEns	55.9	50.7	50.7	50.0	89.8
MDSEns*	55.9	1.2	1.2	50.0	1.9
RMDS	87.0	81.2	81.2	80.6	89.8
Gram	-	-	-	-	-
EBO	82.2	77.4	77.4	76.8	89.8
GradNorm	43.8	37.1	37.1	50.2	89.8
ReAct	81.8	76.6	76.6	60.8	89.8
ReAct*	81.8	19.5	19.5	54.6	21.9
MLS	83.2	78.4	78.4	77.7	89.8
KLM	-	-	-	-	-
VIM	81.8	76.9	76.9	74.8	89.8
KNN	84.7	78.7	78.7	77.3	89.8
DICE	78.7	73.8	73.8	50.0	89.8
DICE*	75.0	7.8	7.8	51.3	8.8
RankFeat*	65.2	22.4	22.4	50.3	32.0
RankFeat	65.2	60.3	60.3	50.0	89.8
ASH	82.1	76.3	76.3	76.3	87.8
SHE	75.5	69.8	69.8	68.4	89.8
NECO	-	-	-	-	-
FDBD	89.6	83.9	83.9	83.2	89.8
RP$_{\text{MSP}}$	91.7	85.1	85.1	84.5	89.8
RP$_{\text{ODIN}}$	88.6	81.6	81.6	82.4	89.7
RP$_{\text{EBO}}$	80.1	75.5	75.5	75.2	89.8
RP$_{\text{GradNorm}}$	31.9	25.7	25.7	50.0	89.8
NCI	83.1	78.3	78.3	77.4	89.8
ConfBranch	77.1	72.8	72.8	72.0	89.8
OpenGAN	31.8	25.8	25.8	43.4	88.6
LogitNorm	87.7	79.7	79.7	81.0	85.5
ARPL	87.7	82.8	82.8	80.8	90.0
GODIN	87.7	82.0	82.0	81.7	90.0
RotPred	86.4	80.0	80.0	79.0	89.7
OE	85.4	77.4	77.4	78.8	85.3
UDG	82.8	71.4	71.4	73.7	80.8
MixOE	91.0	85.3	85.3	84.1	90.3
Energy	91.7	83.9	83.9	82.6	89.6
NovelBranch	89.9	82.5	82.5	80.4	89.9
Extended	87.2	80.2	80.2	77.7	89.8

Table 8: **Full results of W-Europe.** AUROC, AUOSCR, OpenAUC, and OOSA are averaged across three open-set splits: Local, Non-Local, and Non-Moth. For post-hoc methods, we present the result of one of the three runs.

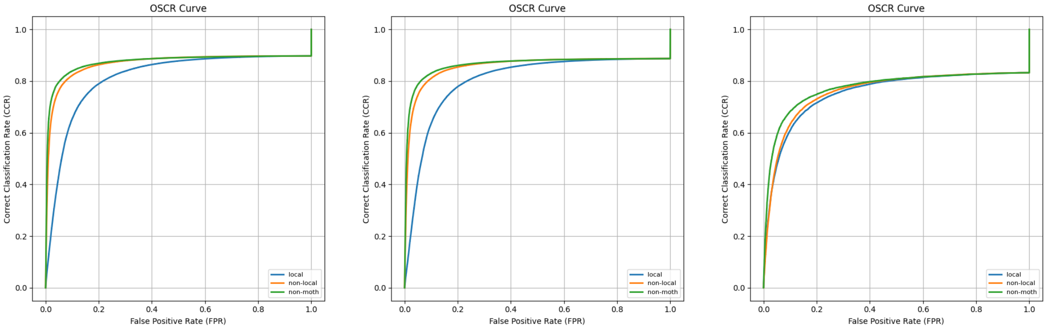
Method	AUROC	AUOSCR	OpenAUC	OOSA	ACC
MSP	91.5	84.0	84.0	83.9	88.7
TempScale	91.6	84.0	84.0	84.0	88.7
ODIN	88.0	80.2	80.2	79.5	88.7
MDS	81.6	76.0	76.0	76.6	88.7
MDSEns	56.3	50.4	50.4	51.7	88.7
MDSEns*	56.3	1.2	1.2	50.1	1.9
RMDS	87.7	80.7	80.7	80.5	88.7
Gram	-	-	-	-	-
EBO	81.3	75.7	75.7	75.6	88.7
GradNorm	41.7	34.5	34.5	50.0	88.7
ReAct	81.9	75.8	75.8	72.4	88.7
ReAct*	81.9	13.6	13.6	53.5	14.8
MLS	82.2	76.6	76.6	76.5	88.7
KLM	-	-	-	-	-
VIM	81.0	75.2	75.2	74.8	88.7
KNN	84.3	77.2	77.2	75.9	88.7
DICE	80.7	74.5	74.5	72.4	88.7
DICE*	80.7	3.0	3.0	50.0	3.2
RankFeat	60.8	55.9	55.9	51.8	88.7
RankFeat*	60.8	30.5	30.5	53.0	41.5
ASH	81.3	75.7	75.7	72.5	88.7
SHE	73.8	66.9	66.9	66.4	88.7
NECO	-	-	-	-	-
FDBD	88.9	82.3	82.3	82.0	88.7
RP_{MSP}	91.5	84.0	84.0	83.9	88.7
RP_{ODIN}	89.7	81.7	81.7	83.5	88.6
RP_{EBO}	79.2	73.9	73.9	74.1	88.7
$\text{RP}_{\text{GradNorm}}$	33.1	26.4	26.4	50.0	88.7
NCI	82.2	76.5	76.5	76.2	88.7
ConfBranch	71.7	67.3	67.3	67.4	88.6
OpenGAN	45.9	40.8	40.8	47.1	87.9
LogitNorm	88.0	78.9	78.9	80.6	84.5
ARPL	86.9	81.1	81.1	79.2	89.0
GODIN	88.7	81.7	81.7	81.5	88.8
RotPred	85.3	77.6	77.6	77.0	87.9
OE	83.2	73.0	73.0	74.6	83.6
UDG	-	-	-	-	-
MixOE	90.4	83.8	83.8	82.9	89.1
Energy	89.7	81.0	81.0	79.7	88.4
NovelBranch	89.8	81.6	81.6	80.5	88.6
Extended	88.0	80.0	80.0	78.5	88.6

Table 9: **Full results of C-America.** AUROC, AUOSCR, OpenAUC, and OOSA are averaged across three open-set splits: Local, Non-Local, and Non-Moth. For post-hoc methods, we present the result of one of the three runs.

Method	AUROC	AUOSCR	OpenAUC	OOSA	ACC
MSP	87.6	79.5	79.5	80.6	85.0
TempScale	85.6	76.3	76.3	77.8	83.2
ODIN	86.4	78.1	78.1	42.5	85.0
MDS	85.8	77.9	77.9	79.2	84.6
MDSEns	62.6	54.5	54.5	55.8	85.0
MDSEns*	62.6	9.3	9.3	50.6	11.7
RMDS	85.6	76.2	76.2	78.6	83.2
Gram	50.9	43.1	43.1	50.0	85.0
EBO	87.5	79.6	79.6	81.4	85.0
GradNorm	28.9	21.0	21.0	50.0	85.0
ReAct	86.2	78.0	78.0	50.8	84.5
MLS	88.3	80.2	80.2	81.5	85.0
KLM	85.1	77.5	77.5	79.4	85.0
VIM	86.5	78.8	78.8	79.7	85.0
KNN	84.4	76.6	76.6	77.6	85.0
DICE	20.8	13.2	13.2	50.0	85.0
DICE*	20.8	12.1	12.1	50.0	82.7
RankFeat	76.6	69.7	69.7	69.0	85.0
RankFeat*	76.6	53.9	53.9	65.9	60.4
ASH	87.5	79.6	79.6	81.4	85.0
SHE	82.6	75.1	75.1	74.1	85.0
NECO	65.4	57.5	57.5	42.4	84.7
FDBD	88.1	80.0	80.0	81.0	85.0
$\mathrm{RP}_{\mathrm{MSP}}$	87.7	79.6	79.6	80.6	85.0
$\mathrm{RP}_{\mathrm{ODIN}}$	87.5	79.2	79.2	42.5	85.0
$\mathrm{RP}_{\mathrm{EBO}}$	88.2	80.4	80.4	82.1	85.0
$\mathrm{RP}_{\mathrm{GradNorm}}$	18.7	11.2	11.2	50.0	85.0
NCI	87.7	79.7	79.7	81.0	85.0
ConfBranch	84.2	77.1	77.1	77.9	85.5
OpenGAN	45.1	35.5	35.5	50.0	82.2
LogitNorm	89.2	81.0	81.0	82.3	85.7
GODIN	72.8	63.7	63.7	63.4	83.1
RotPred	86.1	78.2	78.2	79.4	85.3
OE	91.9	80.5	80.5	83.0	84.0
UDG	83.9	70.1	70.1	75.3	76.0
MixOE	88.1	79.7	79.7	81.4	84.9
Energy	91.7	80.4	80.4	82.9	83.8
NovelBranch	89.5	81.0	81.0	82.5	85.6
Extended	88.4	80.1	80.1	81.0	85.8

Table 10: **Benchmarking results on Open-Insect for C-America with ImageNet-1K pretrained weights.** We evaluate approaches falling into three categories: i) post-hoc methods, ii) training-time regularization (without auxiliary data), and iii) outlier exposure with auxiliary data. Results are shown for the C-America. For each of the three open-set splits – local (L), non-local (NL), and non-moth (NM) – the AUROC is reported along with the accuracy of the closed-set test set. Since C-America is the dataset with the smallest training dataset, we also report results using ImageNet-1K pretrained weights compared to the models trained from scratch in Table 2.

	C-America		
	L	NL	closed-set Acc.
Post-hoc methods			
OpenMax	N/A	N/A	**89.14**
MSP	86.01	87.02	
TempScale	85.42	86.54	
ODIN	69.83	70.72	
MDS	86.95	88.39	
MDSEns	57.1	59.38	
RMDS	**89.0**	**91.35**	
Gram	38.06	42.7	
EBO	86.48	88.72	
OpenGAN	64.41	68.17	
GradNorm	37.05	35.6	
ReAct	86.62	88.79	
MLS	87.61	89.67	
KLM	87.36	88.5	
VIM	88.05	89.74	
KNN	86.3	88.01	
DICE	19.12	16.5	
RankFeat	68.33	69.27	
ASH	85.19	86.67	
SHE	57.55	60.8	
NECO	80.87	81.76	
FDBD	87.76	88.97	
RP-MSP	86.0	87.15	
RP-ODIN	70.28	73.75	
RP-EBO	85.35	88.09	
RP-GradNorm	27.62	25.73	
Training-time regularization			
ConfBranch	81.38	93.23	89.69
LogitNorm	**88.41**	**95.38**	**89.86**
ARPL	87.99	92.17	89.29
G-ODIN	71.02	65.93	88.42
RotPred	80.28	92.04	89.28
Training with auxiliary data			
OE	89.16	93.32	88.71
MCD	88.54	91.38	88.38
UDG	82.27	88.97	78.91
MIXOE	90.07	94.24	89.12
Energy	90.64	93.66	88.0
NovelBranch	**92.05**	**95.35**	89.83
Extended	91.19	93.75	**90.67**



(a) NE-America (b) W-Europe (c) C-America

Figure 6: **OSCR curves of the basic classifier with MSP as the post-hoc method across three regions.** The x-axis is the False Positive Rate (FPR) and the y-axis is the Correct Classification Rate (CCR).

Table 11: **Benchmarking results on Open-Insect with different training and post-hoc method combinations.** We present the results obtained from different combinations of training methods (with or without auxiliary training data) with selected post-hoc methods. Results are shown for the three regions in Open-Insect: NE-America, W-Europe, C-America. For each of the three open-set splits – local (L), non-local (NL), and non-moth (NM) – the AUROC is reported. The colored AUROC are those shown in Table 2.

		NE-America			W-Europe			C-America		
Training	Post-hoc	L	NL	NM	L	NL	NM	L	NL	NM
Training regularization										
LogitNorm	MSP	77.45	84.28	94.0	76.87	85.19	88.7	87.35	89.12	90.22
	EBO	80.55	87.25	95.32	80.77	87.72	95.61	87.3	89.54	90.01
	MLS	80.58	87.29	95.31	80.78	87.73	95.59	87.27	89.44	90.15
	TempScale	79.96	86.5	94.99	79.2	86.06	94.87	87.56	89.5	90.47
Training with auxiliary data										
OE	MSP	79.75	86.33	90.12	75.36	84.58	89.6	89.54	94.03	92.16
	MLS	56.43	52.96	54.37	53.68	47.23	44.7	84.58	86.13	88.28
	EBO	55.82	52.05	53.1	50.71	42.65	39.19	73.26	68.79	75.8
	TempScale	79.48	86.36	90.14	75.38	84.58	89.61	89.51	94.04	92.14
UDG	MSP	74.11	80.22	92.56	-	-	-	80.76	83.08	87.89
	MLS	75.1	81.94	91.48	-	-	-	80.36	81.75	85.9
	EBO	71.73	77.45	75.45	-	-	-	72.1	70.65	70.9
	TempScale	74.63	80.19	92.1	-	-	-	80.51	82.31	87.34
MixOE	MSP	86.16	92.45	94.39	85.21	91.93	94.1	86.19	87.94	90.12
	MLS	72.43	78.56	90.61	69.34	75.95	90.04	85.51	87.06	91.03
	EBO	72.03	78.01	90.29	68.9	75.35	89.73	84.91	86.36	90.57
	TempScale	86.99	93.99	94.81	85.7	93.58	94.64	86.04	88.03	89.94
Energy	MSP	81.86	86.97	93.58	80.81	84.46	93.3	86.32	88.33	90.13
	MLS	87.37	95.12	92.53	84.64	94.79	89.53	89.91	93.62	91.49
	EBO	87.21	94.98	91.6	84.34	94.6	88.05	89.99	93.79	91.22
	TempScale	82.1	87.53	94.82	81.04	85.04	94.76	87.39	89.76	91.56
NovelBranch	MSP	82.5	87.74	93.5	81.16	86.15	93.02	86.54	88.08	90.47
	MLS	85.51	94.07	89.98	83.94	93.76	91.74	87.77	89.65	91.10
	EBO	85.31	93.92	88.96	83.67	93.63	90.82	87.67	89.54	90.82
	TempScale	83.17	88.86	95.19	81.94	87.79	95.06	87.20	88.95	91.84
Extended	MSP	82.4	87.76	93.39	80.98	85.89	92.82	85.78	86.88	90.34
	MLS	83.49	92.15	86.08	82.63	91.94	89.38	86.91	89.03	89.31
	EBO	83.25	91.94	84.86	82.33	91.75	88.29	86.81	88.94	88.95
	TempScale	83.09	88.89	95.11	81.78	87.48	94.89	86.37	87.55	91.81

Table 12: AUROC scores for post-hoc methods with high standard deviation across three training runs. We highlight the AUROC in red when the OSR scores for positive (open-set) samples are generally lower than those for negative (closed-set) samples.

Method	L				NL				NM			
	Run 1	Run 2	Run 3	Mean$_{\text{(standard deviation)}}$	Run 1	Run 2	Run 3	Mean$_{\text{(standard deviation)}}$	Run 1	Run 2	Run 3	Mean$_{\text{(standard deviation)}}$
NE-America												
KNN [78]	76.61	39.49	76.90	64.33 (17.57)	84.43	38.24	84.32	68.99 (21.75)	95.83	20.41	94.80	70.35 (35.32)
DICE [77]	67.47	42.77	67.38	59.21 (11.62)	74.46	41.15	73.11	62.91 (15.39)	94.32	37.38	92.21	74.64 (26.36)
SHE [101]	68.56	38.17	68.67	58.47 (14.35)	76.23	36.87	75.62	62.91 (18.41)	88.49	24.75	89.19	67.47 (30.22)
W-Europe												
KNN [78]	76.00	51.13	74.79	67.30 (11.45)	85.15	43.34	84.91	71.13 (19.65)	95.20	35.19	93.95	74.78 (28.00)
DICE [77]	68.66	42.54	60.84	57.35 (10.95)	79.13	42.03	65.79	62.32 (15.35)	94.37	48.36	92.51	78.41 (21.26)
SHE [101]	67.13	44.63	63.47	58.41 (9.86)	73.65	39.18	70.75	61.19 (15.61)	87.83	34.27	83.54	68.55 (24.30)

C.3 Comparing the efficiency of post-hoc methods

All methods were evaluated with 1 RTX8000 GPU, 8 CPUs, 8 workers, and 100 GB CPU memory. The classifier used is the ResNet50 C-America classifier. In Table 13, if "Data needed for setup" is "Train", the setup time was calculated with 3,167 images, a subset of C-America training images. If "Data needed for setup" is "Val", the setup time was calculated with 2,000 images, a subset of C-America validation images. Setup only need to be done once and the time needed is independent of the number of test images. Here, the inference time in Table 13 is the time to process 4,000 test images (2,000 closed-set and 2,000 open-set). Inference time increases as the number of test images increases. Some methods require hyperparameters. We indicate this by placing ✓in the "Hyperparameter Search" column of Table 13.

Table 13: **Comparison of post-hoc method efficiency.** We compare the efficiency of post-hoc methods in terms of setup time, access to training data, inference efficiency, and hyperparameter search requirements. Times are reported in seconds.

Method	Setup time (Sec.)	Data needed for setup	Inference time (Sec.)	Hyperparameter search
OpenMax [6]	10.47 (2.47)	Train	34.24 (2.82)	
MSP [31]	N/A	N/A	7.52 (1.98)	
TempScale [26]	5.63 (1.96)	Val	5.16 (0.86)	
ODIN [50]	N/A	N/A	11.05 (2.29)	✓
MDS [49]	17.24 (1.29)	Train	66.92 (7.89)	
MDSEns [49]	10.90 (1.65)	Train	9.21 (0.24)	✓
RMDS [64]	27.12 (1.56)	Train	70.41 (2.22)	
Gram [68]	19.12 (1.64)	Train	17.19 (1.93)	✓
EBO [54]	N/A	N/A	4.92 (0.49)	✓
GradNorm [39]	N/A	N/A	10.91 (1.87)	
ReAct [76]	6.12 (1.67)	Val	5.65 (0.43)	✓
MLS [29]	N/A	N/A	10.74 (1.77)	
KLM [30]	7.71 (2.12)	Val	1180.58 (9.89)	
VIM [87]	22.57 (2.19)	Train	8.73 (2.01)	✓
KNN [78]	11.54 (2.25)	Train	6.85 (0.93)	✓
DICE [77]	10.46 (1.65)	Train	6.12 (0.34)	
RankFeat [73]	N/A	N/A	19.40 (1.87)	
ASH [17]	N/A	N/A	6.10 (1.22)	✓
SHE [101]	10.71 (2.13)	Train	5.87 (0.72)	
NECO [3]	16.85 (3.71)	Train	6.74 (2.02)	✓
FDBD [51]	10.30 (1.55)	Train	5.97 (1.30)	✓
RP$_{\text{MSP}}$ [43]	N/A	Train	9.88 (3.46)	
RP$_{\text{ODIN}}$ [43]	N/A	Train	9.90 (0.30)	✓
RP$_{\text{EBO}}$ [43]	N/A	Train	4.93 (0.42)	✓
RP$_{\text{GradNorm}}$ [43]	N/A	Train	9.94 (3.16)	
NCI [52]	9.53 (1.88)	Train	5.18 (0.75)	✓

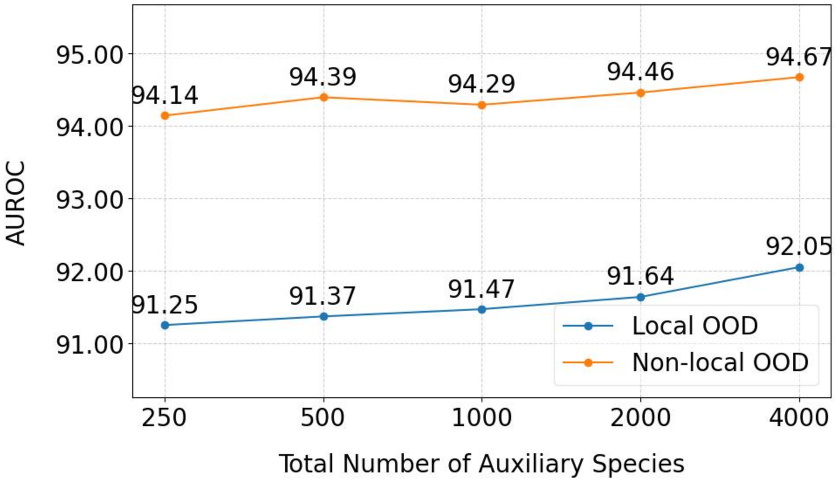


Figure 7: **Species diversity of the auxiliary dataset.** We vary the number of species in the auxiliary dataset, while keeping the total number of images fixed to 80,000.

D Explainability of OSR methods

We constructed the following dataset to empirically verify that *background features are not enough to achieve good OSR performance on Open-Insect*. We first used Grounding DINO [53] to detect and generate a bounding box around the insect in each image, then used it as an input prompt to Segment Anything (SAM) [44], which segmented the insect. Finally, we replaced the segmented region with the average value of the surrounding pixels. We applied this pipeline to a subset of Open-Insect C-America. Since both object detection and segmentation were done by models, errors can occur. Hence, we manually verified a (randomized) subset of the processed images and discarded the rest, ultimately obtaining 859 valid closed-set and 1433 valid local open-set images for our experiments. We show an example in Fig. 1.

We believe the slight drop of performance when the background is masked can be explained by the quality of the machine-generated masks. During verification, we noticed that the masks generated by SAM often miss legs and antennae (see Fig. 9). As a result, when masking the background, parts of the insect such as legs and antennae were also inadvertently masked, which may explain the slight drop in performance compared to the original images.

(a) Original image

(b) Moth masked

(c) Background masked

Figure 8: An example from the subset we used for the explainability experiment. The mask was automatically generated by SAM and the color of the mask is the average value of the surrounding pixels.

E Performance of BioCLIP on Open-Insect

BioCLIP [74], a foundation model for species recognition, is increasingly used for biodiversity-related tasks. Hence, we conducted additional experiments to evaluate the OSR performance obtained directly from BioCLIP as well as models finetuned from BioCLIP vision encoder weights.

(a) Original image (b) Moth masked (c) Background masked

Figure 9: A mask generated by SAM which misses legs and antennae.

E.1 Off-the-shelf performance of BioCLIP on Open-Insect

Since not all closed-set species of Open-Insect are included in BioCLIP's training set and some open-set species overlap with BioCLIP's training data, we evaluated BioCLIP's off-the-shelf performance on a subset of Open-Insect. We include closed-set species that are included in Tree-of-Life 10M (training data of BioCLIP) [74] or "seen" by BioCLIP and open-set species that are "unseen" to evaluate its OSR performance. We list the number of species of each regional split in Table 14. We also include 4,633 "unseen" non-local moth species in this subset.

Table 14: **Number of species in each region of the Open-Insect subset.** Seen closed-set species appear in both the Open-Insect and BioCLIP training sets, while unseen open-set species are open-set species in Open-Insect and do not appear in BioCLIP training set.

Category	NE-America	W-Europe	C-America
Seen closed-set	1260	177	71
Unseen open-set	418	447	1537

We compared BioCLIP's performance on this subset with our ResNet50 classifiers trained from scratch. We used three simple post-hoc methods, MSP, MLS, and EBO. We find that open-set recognition on Open-Insect is challenging for BioCLIP, likely because Lepidoptera (moths and butterflies) represent less than 2% of the species in its training data (Table 15).

Table 15: **Comparison of BioCLIP's performance on the Open-Insect subset with our ResNet50 classifiers trained from scratch.**

Model	Post-hoc Method	NE-America			W-Europe			C-America		
		L	NL	closed-set Acc.	L	NL	closed-set Acc.	L	NL	closed-set Acc.
BioCLIP	MSP	65.24	60.38	24.37	68.83	69.30	37.56	74.83	73.33	55.75
	MLS	77.86	72.96		76.44	79.54		79.64	85.72	
	EBO	55.94	56.19		56.05	57.10		53.08	63.73	
Ours	MSP	87.71	93.97	90.49	85.61	92.74	89.07	83.89	84.52	78.52
	MLS	75.25	76.96		69.64	75.49		74.63	74.61	
	EBO	74.60	75.93		68.99	74.57		71.89	71.71	

E.2 Using BioCLIP pretrained weights

It has been shown that using ImageNet pretrained weights can help closed-set accuracy [81] and that there is a positive correlation between the closed-set and open-set performance [84]. The results we show in Table 10 show that those observations hold for Open-Insect as well. Since BioCLIP was trained on more domain specific data, we finetuned two ViT-B-16 models [18] for C-America - one with BioCLIP pretrained weights, and the other with ImageNet-1K pretrained weights, to compare the effect of different pretraining data.

Both models were fine-tuned for 30 epochs with all 636 C-America closed-set species with 224×224 images. We only finetuned the last block and kept other parameters frozen. The OSR performance was evaluated with all 636 closed-set species, 1537 "unseen" local (L) open-set, and 4,633 "unseen" non-local (NL) open-set species. We present the results in Table 3. We find that using BioCLIP weights is more effective than ImageNet-1K.