From Coarse to Fine-Grained Open-Set Recognition

Supplementary Material

A. Details on hierarchy-adversarial learning

This section provides further details on our hierarchy-adversarial learning approach introduced in section 3.3. We are interested in learning an encoder $f_{\boldsymbol{\theta}}$ that yields a representation that can be used to classify fine-grained categories – but is agnostic to coarser hierarchical labels at level $t > 0$ in the hierarchy. For completeness, we repeat Eq. 5 from the main paper and recall that $p^t(y_i^t|\boldsymbol{x}_i) = p(y_i^t|\boldsymbol{x}_i; \boldsymbol{\theta}, \boldsymbol{\omega}^t)$ at the t-th hierarchy level:

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

Directly minimizing Eq. 5 is problematic, as we allow both adjustments of the encoder $f_{\boldsymbol{\theta}}$ and the linear decoding functions $h_{\boldsymbol{\omega}^t}$ (with $t > 0$) to maximize the cross-entropy of the coarse granularities $\mathcal{L}^t$. First, the terms $\mathcal{L}^t$ can be maximized mainly by $h_{\boldsymbol{\omega}^t}$, leaving the encoder and the resulting representations unaffected. Second, $h_{\boldsymbol{\omega}^t}$ can learn trivial solutions, e.g., by recognizing one of the easy categories and always predicting zero probability for that category if the input belongs to that category and a probability of one otherwise.

To guarantee that the hierarchy-adversarial objective affects the encoder, we formulate the optimization problem as a min-max optimization problem where λ controls the trade-off between the fine-grained classification and the *regularization* through hierarchy indifference. We denote the parameters of the linear decoder $h_{\boldsymbol{\omega}^t}$ as $\boldsymbol{\omega}^t$, which includes the weights and bias. We can rewrite the objective as a function of $\boldsymbol{\theta}$ and $\boldsymbol{\omega}^t$ where $\mathcal{L}^t$ corresponds to the cross-entropy (negative log-likelihood) at level t and $\mathcal{L}^0$ to the cross-entropy for the fine-grained labels:

$$\mathcal{L}_{\text{HA}}(\boldsymbol{\theta}, \boldsymbol{\omega}^0, \boldsymbol{\omega}^1, \ldots, \boldsymbol{\omega}^T) = \mathcal{L}^0(\boldsymbol{\theta}, \boldsymbol{\omega}^0) - \lambda \sum_{t=1}^{T} \mathcal{L}^t(\boldsymbol{\theta}, \boldsymbol{\omega}^t). \tag{1}$$

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

$$\min_{\boldsymbol{\theta}, \boldsymbol{\omega}^0} \max_{\boldsymbol{\omega}^1, \ldots, \boldsymbol{\omega}^T} \mathcal{L}_{\text{HA}}(\boldsymbol{\theta}, \boldsymbol{\omega}^0, \boldsymbol{\omega}^1, \ldots, \boldsymbol{\omega}^T). \tag{2}$$

That is, we obtain the parameters $\boldsymbol{\theta}$ of the encoder that *minimize* $\mathcal{L}^0$ (i.e., the representation is discriminative for fine-grained labels) and at the same time *maximize* $\mathcal{L}^t$, with $t > 0$, that is, the representation is pushed towards not capturing coarser-grained hierarchical labels.

To train a deep neural network with this min-max objective using stochastic gradient descent (SGD), we adapt a *gradient reversal layer*-based approach inspired by the domain adaptation literature [2, 3]. By swapping the sign of the gradient corresponding to the terms $\mathcal{L}^t$ before back-propagating into the encoder (see illustration in Fig. 2c in the main paper), this allows us to update the learnable parameters using standard SGD with one forward- and backward-pass.

B. iNat2021-OSR dataset

B.1. Open-set splitting protocol

Here, we provide additional information about how we constructed the datasets used for evaluation in the main paper. First, we select a subset of the iNat2021 dataset [5], based on a super-category at a certain hierarchy level, as the training domain. Specifically, we use Aves (birds) and Insecta (insects) at the "Class" level, as they are the dominant super-categories in iNat2021 and provide the largest number of images. We then select a subset of the fine-grained categories within these super-categories as the closed-set (familiar) categories used for training. We obtain seven distinct open-sets with semantic distances ranging from 1-hop to 7-hop, the hop measuring the taxonomic distance between an open-set category and its *nearest* training category.

We achieve this in an iterative way, starting at the root node T to select open-set candidates O^t that preserve a large number of training categories and then guarantee that every open-set category has the nearest training category at semantic distance t-hop.

For every level t of the hierarchy:

1. We start by setting the open-set candidates $O^t = \mathcal{Y}^t$.
2. Exclude all open-set categories already selected at coarser hierarchy levels from O^t.
3. Exclude 20% of categories with the highest number of samples from O^t.
4. Exclude one "sibling" (category with same parent) of every category y^t in O^t to guarantee that every remaining open-set candidate has the nearest training category at semantic distance t-hop.
5. Select the final t-hop open-set by randomly selecting $0.2|\mathcal{Y}^t|$ categories of O^t and use their corresponding fine-grained categories y^0 as the final t-hop open-set. In cases where $|O^t| < 0.2|\mathcal{Y}^t|$ we select $0.5|O^t|$ categories.

Categories not assigned to any of the final t-hop open-sets are used as the closed-set categories. Given this closed-

	Train	Test	Open-set test						
	familiar	familiar	1-hop	2-hop	3-hop	4-hop	5-hop	6-hop	7-hop
Kingdom	1	1	1	1	1	1	1	1	2
Phylum	1	1	1	1	1	1	1	5	7
Class	1	1	1	1	1	1	6	20	24
Order	26	26	13	15	10	6	48	72	121
Family	105	105	48	41	29	13	202	366	388
Genus	379	379	118	129	86	54	539	1,801	1,909
Species	745	745	297	180	170	94	930	2,972	4,607
Samples	210,323	7,450	2,970	1,800	1,700	940	9,300	29,720	46,070

Table A1. **iNat2021-OSR-Aves data split statistics.**

	Train	Test	Open-set test						
	familiar	familiar	1-hop	2-hop	3-hop	4-hop	5-hop	6-hop	7-hop
Kingdom	1	1	1	1	1	1	1	1	2
Phylum	1	1	1	1	1	1	1	5	7
Class	1	1	1	1	1	1	6	20	24
Order	14	14	8	8	9	3	16	119	120
Family	129	129	58	57	38	23	68	457	387
Genus	1,039	1,039	248	294	75	64	149	1,367	1,904
Species	1,501	1,501	505	333	99	88	226	2,636	4,587
Samples	398,952	15,010	5,050	3,330	990	880	2,260	26,360	45,870

Table A2. **iNat2021-OSR-Insecta data split statistics.**

set, we compute the semantic distance to categories outside the super-category (i.e., other animals, plants, fungi, etc.) and group them by their hop distance. In both cases of Aves and Insecta, hop distances 5–7 are outside the super-category and thus outside the training domain. The final data split statistics for *iNat2021-OSR-Aves* and *iNat2021-OSR-Insecta* are given in Tables A1 and A2.

B.2. Qualitative examples

We show examples of open-set samples at different hop distances in Fig. A1 and A2.

C. Evaluation

Closed-set. We report overall *accuracy* as well as *macro accuracy* by averaging category accuracies. The latter is relevant for class-imbalanced datasets, which is the case for coarser granularities in iNat2021.

Open-set. To evaluate the binary classification task familiar *vs*. novel, we define the familiar as positive and novel as negative categories. We use standard open-set evaluation metrics to evaluate the ranking of open-set vs. closed-set. These metrics are threshold-free, i.e., the decision threshold is swept over the entire range of possible values of a score to report the average performance over all possible settings [1, 6] and include the *AUROC*: the area under the Receiver-Operator Curve, which plots the true positive rate (TPR) vs. the false positive rate (FPR), and the *AUPR*: area under the Precision-Recall-Curve. Furthermore, we report the *TNR@TPR95*: the true negative rate (true novelty rate) at the true positive rate (true familiarity rate) of 95%. This metric reflects a scenario, where we would calibrate a system to have a high TPR (recall), but would like to optimize the TNR (or false positive rate). Finally, the *Top-5-acc* measures the accuracy of correctly detected novelties in the top 5% retrieved novelties, i.e., the analog to top-k accuracy, but with k set as a percentage of the number of test samples (top-p). This can be critical in discovery applications, where a human expert checks the retrieved results.

D. Additional results

D.1. Closed-set accuracy for hierarchy-aware strategies

Closed-set performance for hierarchy-aware training strategies and the standard cross-entropy training on fine-grained categories are presented in Table A3. The hierarchy-supportive approach improves closed-set accuracy for coarser granularities (genus, family, order), but not for fine-grained categories (species). The hierarchy-adversarial approach has little impact on closed-set accuracy across all granularities.

D.2. Qualitative error cases

We present qualitative error cases in Fig. A7. These false familiarity mistakes from the 1-hop open-set in Aves are the samples with the highest maximum logit score (indicating *familiarity*) when training with cross-entropy on fine-grained categories. We report the confidence of the models trained with cross-entropy, hierarchy-supportive, and hierarchy-adversarial strategies.

D.3. OSR results for hierarchy-aware strategies for NN and KLD scores

For completeness, we include extended results of hierarchy-aware learning strategies for the maximum logit score (MLS) in Fig. A3 (analogous to Fig. 5 in the main paper). Similarly, for the nearest neighbour (NN) score (Fig. A4) and the KL-disagreement (KLD) score (Fig. A5). For Insecta, we observe a clear pattern for the NN score whereby hierarchy-adversarial can improve fine-grained OSR (1-hop), while being harmful for coarse-grained OSR (7-hop). In contrast, hierarchy-supportive training can improve coarse-grained OSR but has little effect on fine-grained OSR. The KL-disagreement score is not improved with hierarchy-aware learning strategies.

D.4. Ablation study for the nearest neighbour (NN) score

In Fig. A6, we present ablation studies for the nearest neighbour (NN) score to analyze the effect of i) normalizing the representations before computing the distance (Fig. A6a,d) and ii) the number of nearest neighbours k used to compute the score (Fig. A6b,c,e,f). First, we observe that in all cases, normalizing the representations leads to a substantial improvement in AUROC (as reported for OOD detection [4]), except for large hop distances (coarse OSR). We

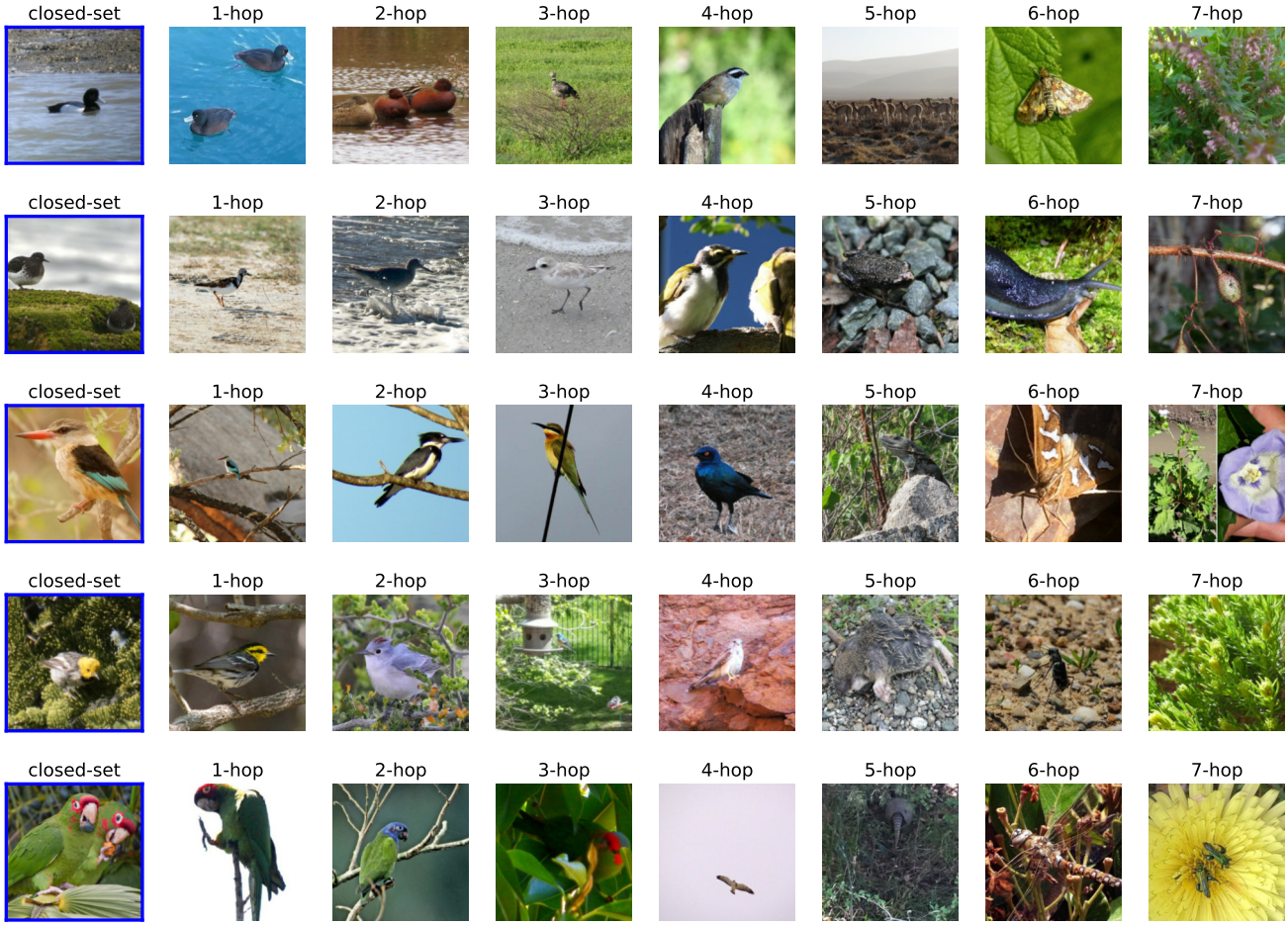

Figure A1. **Examples of semantic distances for iNat2021-OSR-Aves.** Given an image from a closed-set category (blue box), we visualize a random sample from an open-set category with different semantic distance in every row. The taxonomic distance (a proxy for semantic distance) is measured in number of hops, where 1-hop means that two categories share their parent, 2-hop share their grandparent etc.

	Species (fine)	Genus	Family	Order (coarse)
Cross-entropy	67.4 (0.3)	74.9 (0.3)	81.7 (0.3)	89.8 (0.2)
Hierarchy-supportive	66.6 (0.4)	75.8 (0.4)	83.2 (0.4)	91.1 (0.2)
Hierarchy-adversarial (α=0.25)	67.6 (0.3)	75.2 (0.3)	82.2 (0.2)	90.6 (0.2)

(a) Accuracy

	Species (fine)	Genus	Family	Order (coarse)
Cross-entropy	67.4 (0.3)	73.7 (0.3)	78.3 (0.4)	85.1 (0.4)
Hierarchy-supportive	66.6 (0.4)	74.6 (0.4)	79.3 (0.7)	86.2 (0.2)
Hierarchy-adversarial (α=0.25)	67.6 (0.3)	74.1 (0.3)	78.2 (0.7)	86.1 (0.4)

(b) Macro accuracy

Table A3. **Closed-set performance of hierarchy-aware strategies.** Results for single models averaged over 5 independent training runs with standard deviation in parentheses. (a) Overall accuracy. (b) Macro accuracy by averaging the per-category accuracies. The accuracy of coarser test granularities (columns) is evaluated using hard pooling based on the predicted label and the given hierarchy.

note that computing the L2-distance on L2-normalized representations leads to the same ranking as using the cosine-distance, a function of the angle between two feature vectors. Second, increasing the number of nearest neighbours k used to compute the score leads to worse OSR performance, and setting $k = 1$ leads to the best results, except for the 7-hop open-set. This behaviour holds for both using the *mean* and the *max* (i.e., k-th) distance over k nearest neighbours.

References

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

[2] Yaroslav Ganin and Victor Lempitsky. Unsupervised domain adaptation by backpropagation. In *International Conference on Machine Learning (ICML)*, 2015. 1

[3] Yaroslav Ganin, Evgeniya Ustinova, Hana Ajakan, Pascal Germain, Hugo Larochelle, François Laviolette, Mario Marc-

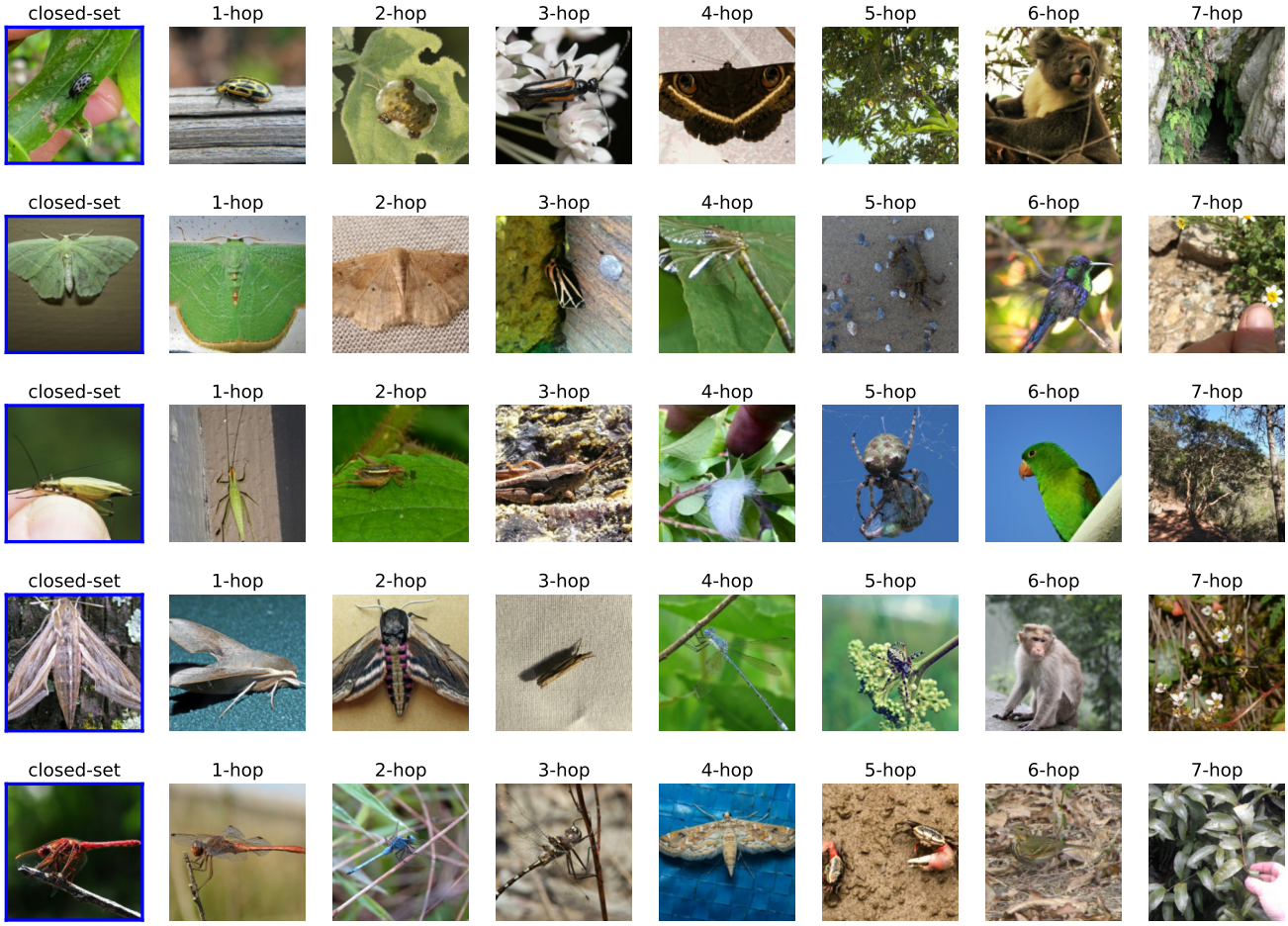

Figure A2. **Examples of semantic distances for iNat2021-OSR-Insecta.** Given a closed-set category (blue box), we visualize a random sample of an open-set category per semantic distance in every row. The taxonomic distance (proxy for semantic distance) is measured in number of hops, where 1-hop means that two categories share their parent, 2-hop share their grandparent etc.

hand, and Victor Lempitsky. Domain-adversarial training of neural networks. *Journal of Machine Learning Research*, 17, 2016. 1

[4] Yiyou Sun, Yifei Ming, Xiaojin Zhu, and Yixuan Li. Out-of-distribution detection with deep nearest neighbors. In *International Conference on Machine Learning (ICML)*. PMLR, 2022. 2

[5] Grant Van Horn, Elijah Cole, Sara Beery, Kimberly Wilber, Serge Belongie, and Oisin Mac Aodha. Benchmarking representation learning for natural world image collections. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021. 1

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

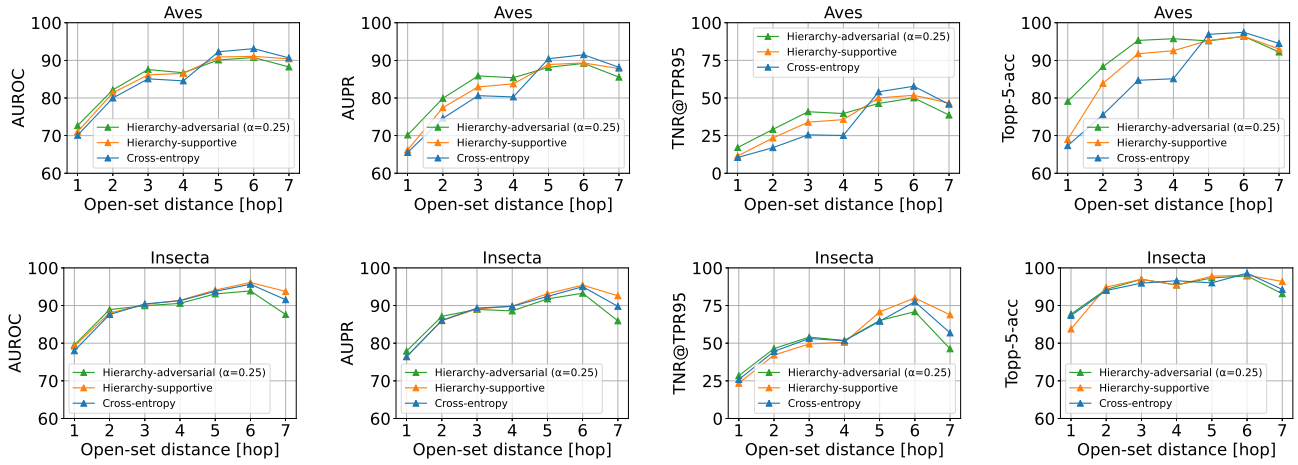

Figure A3. **Hierarchy-aware training strategies for MLS.** OSR results for Aves (top) and Insecta (bottom) using the maximum logit score (MLS) of ensembles with 5 models. We compare the training strategies: Cross-entropy on the fine-grained labels, hierarchy-supportive (Eq. 4, main paper) and hierarchy-adversarial (Eq. 5, main paper). Higher is better for all metrics. Additional evaluation metrics to extend Fig. 5 in the main paper.

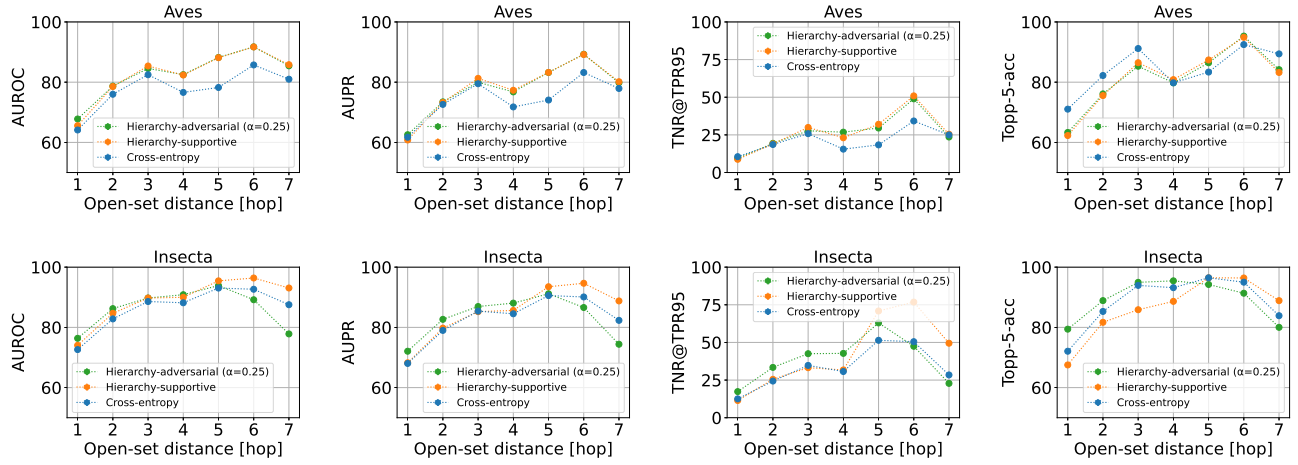

Figure A4. **Hierarchy-aware training strategies for nearest neighbour (NN) score.** OSR results for Aves (top) and Insecta (bottom) using the nearest neighbour (NN) score of ensembles with 5 models. We compare the training strategies: cross-entropy on the fine-grained labels, hierarchy-supportive (Eq. 4, main paper), and hierarchy-adversarial (Eq. 5, main paper). Higher is better for all metrics.

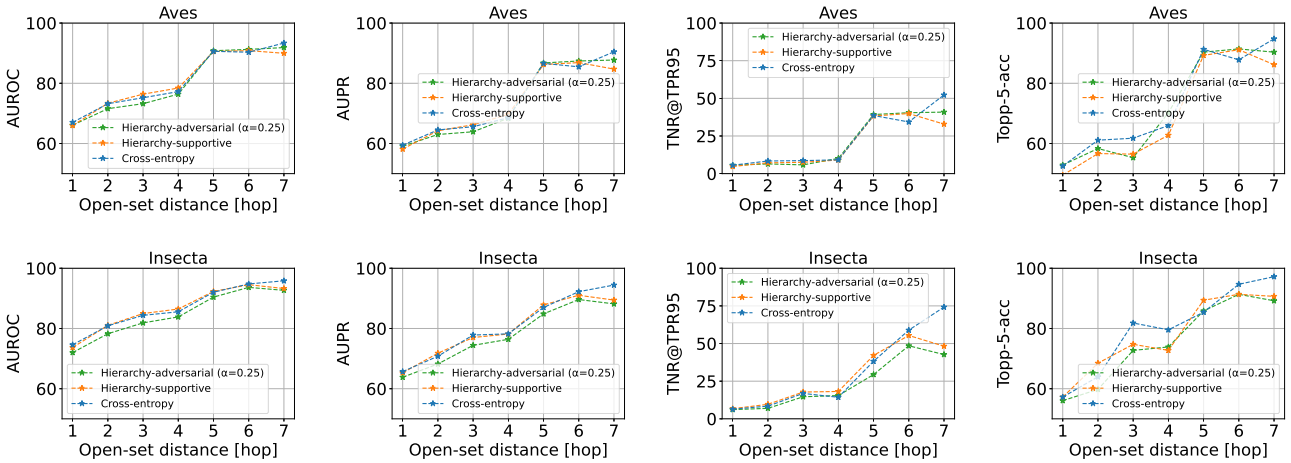

Figure A5. **Hierarchy-aware training strategies for KL-disagreement (KLD) score.** OSR results for Aves (top) and Insecta (bottom) using the Kullback-Leibler disagreement (KLD) score of ensembles with 5 models. We compare the training strategies: cross-entropy on the fine-grained labels, hierarchy-supportive (Eq. 4, main paper), and hierarchy-adversarial (Eq. 5, main paper). Higher is better for all metrics.

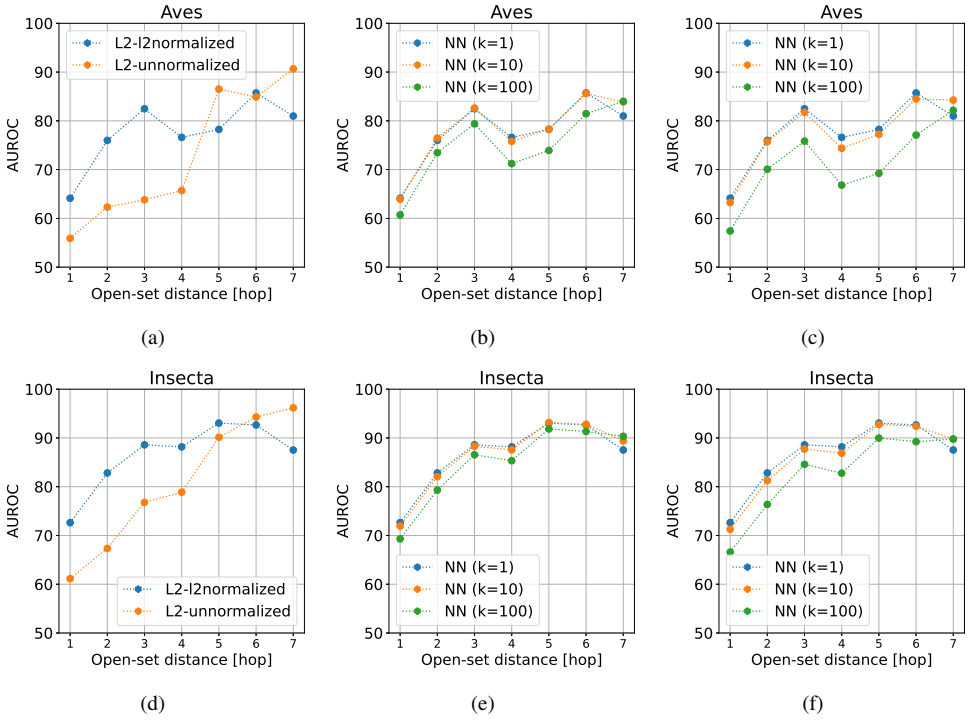

Figure A6. **Ablation study for nearest neighbour (NN) score.** OSR results (AUROC) for Aves (top) and Insecta (bottom). (a,d) The effect of l2-normalizing the representations before computing the L2-distance. (b,e) Number of k nearest neighbours using the *mean* distance. (c,f) Number of k nearest neighbours using the k-th distance, i.e., the *max* distance.

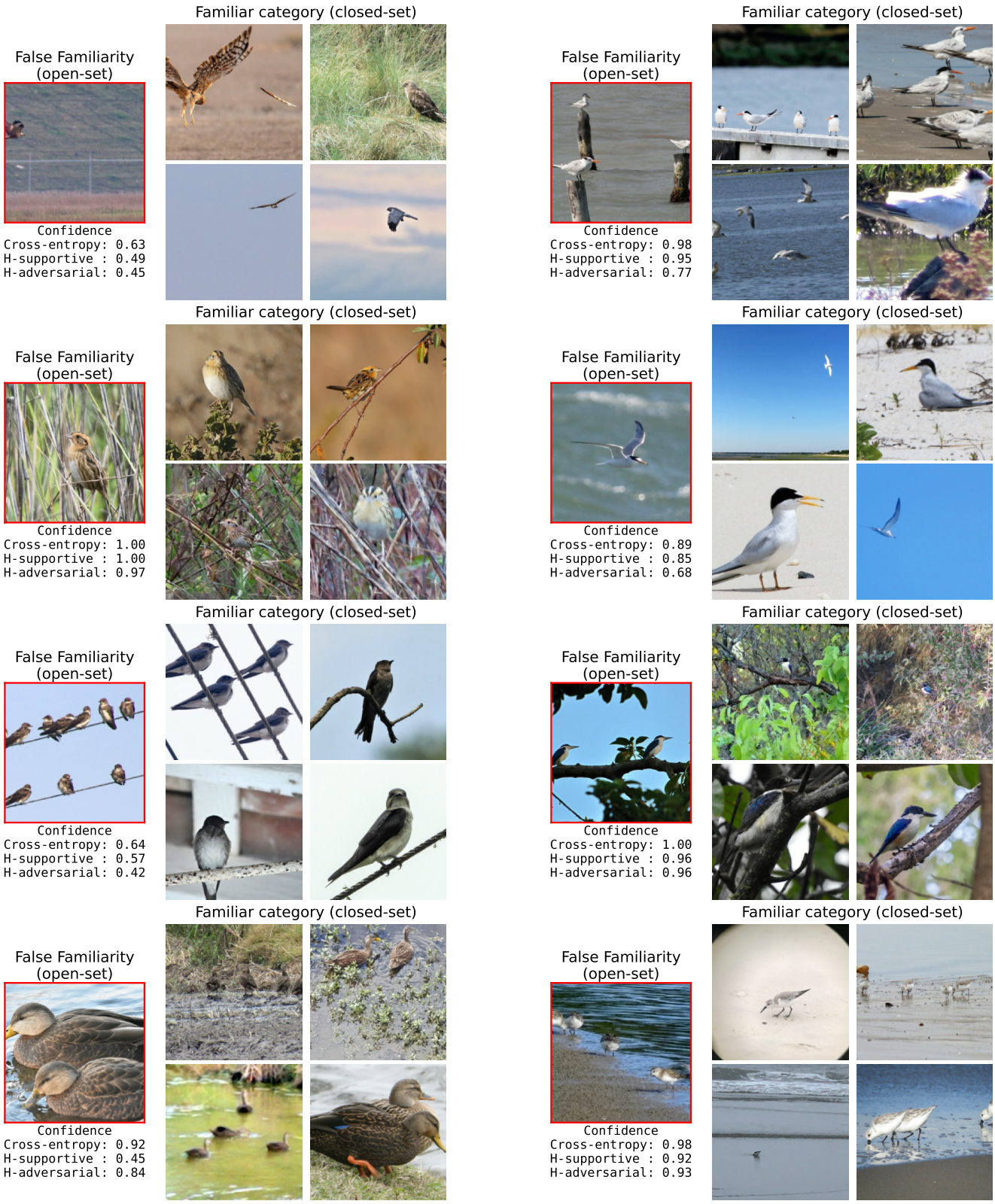

Figure A7. **Error cases.** False familiarity mistakes (i.e., open-set samples confused with closed-set categories with high confidence) of the classifier trained with cross-entropy on the fine-grained labels. The hierarchy-adversarial (H-adversarial) learning reduces the confidence in most of these error cases. All examples are from the 1-hop open-set (red boxes) and are confused with a closed-set category with a 1-hop semantic distance (samples on the right). For non-experts, the visual differences are hard to identify.