PRESERVING FINE-GRAINED STRUCTURE UNDER COARSE SUPERVISION VIA EMA-GUIDED SIMILARITY-WEIGHTED CONTRASTIVE LEARNING

Anonymous authors
Paper under double-blind review

ABSTRACT

Fine-tuning a pretrained model with coarse labels can destroy existing fine-grained structure in latent space. Preserving this latent structure is important when detailed annotations are unavailable. Standard supervised contrastive learning (SupCon) amplifies this problem by attracting all samples sharing a coarse label uniformly, despite potentially substantial within-class heterogeneity. We introduce **TaxoCon**, a similarity-weighted extension of SupCon that uses a stop-gradient exponential moving average (EMA) teacher to estimate within-class pairwise similarity and selectively redistribute positive-pair attraction. Similar positives therefore receive greater learning pressure, while dissimilar positives are attracted less strongly, without requiring fine-grained labels, an auxiliary self-supervised objective, or an external memory bank. We evaluate TaxoCon on iNaturalist 2021 and FGVC-Aircraft using only coarse labels for representation fine-tuning and model selection, with finer-grained labels withheld for post-hoc evaluation. Across ResNet-50, ViT-S/16, and Swin-T pretrained backbones, TaxoCon consistently improves preservation of fine-grained structure across neighborhood retrieval, linear-probing, and hierarchy-sensitive metrics on evaluated baselines. These results show that TaxoCon can mitigate the loss of fine-grained structure during coarse-supervised fine-tuning.

1 INTRODUCTION

Pretrained visual representations often contain structure at multiple levels of semantic granularity Huang et al. (2021); Olson et al. (2025). However, when such representations are fine-tuned using only coarse labels, their pre-existing fine-grained structure can be degraded by the training objective. This is particularly relevant in settings where fine-grained annotations require substantial expertise, time, and annotation effort, while coarse labels are considerably more accessible.

Fine-grained recognition in biodiversity monitoring provides a natural example of this setting. Datasets such as iNaturalist 2021 Van Horn et al. (2021) epitomize this challenge: they contain images spanning thousands of species, organized not as a flat label set but a rich taxonomic hierarchy—*genus* within *families*, *families* within *orders*, and so on. This hierarchy reflects meaningful morphological and ecological relationships between categories. Figure 1 illustrates a taxonomic tree within a particular *Order*. Obtaining labels at finer ranks of such hierarchies can require greater domain expertise, time, and annotation effort, making fine-grained annotation difficult to scale Cui et al. (2018). This motivates a practical question: can a model fine-tuned using only coarse labels preserve the finer-grained structure already present in its pretrained representation?

In particular, methods such as supervised contrastive learning (SupCon) treat all samples sharing a coarse label as equally valid positives, encouraging them to become uniformly close in the representation space despite potentially substantial visual differences. We investigate whether this problem can be mitigated without introducing finer-grained supervision. We propose **TaxoCon**, which replaces uniform positive-pair treatment with EMA-guided similarity weighting, assigning stronger attraction to visually similar samples within a coarse class and weaker attraction to dissimilar ones. Figure 2 illustrates this intuition. TaxoCon requires no fine-grained labels, instead using similarity in the evolving representation as an implicit signal.

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

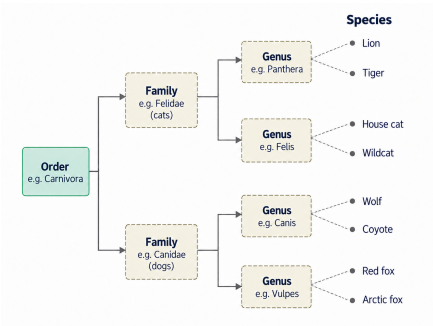

Figure 1: Taxonomic hierarchy where a coarse *order* groups diverse organisms and finer ranks distinguish more specific categories.

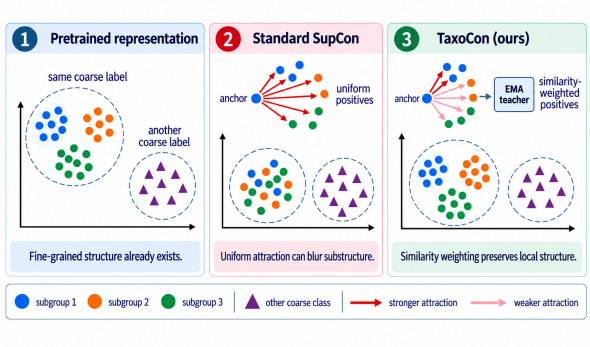

Figure 2: SupCon applies uniform attraction across same-coarse-label samples, whereas TaxoCon redistributes this attraction using EMA-estimated similarity, reinforcing local structure without access to fine-grained labels.

Our contributions are as follows:

- We formulate coarse-supervised fine-tuning as a representation-preservation problem. Rather than asking how to discover new fine-grained classes from coarse labels, we study how to prevent fine-tuning from degrading fine-grained structure that is already encoded in a pretrained representation.
- We propose **TaxoCon**, a contrastive objective specifically designed for this preservation setting. Unlike other coarse-to-fine approaches that seek to discover or impose finer-grained structure through pseudo-labeling, auxiliary objectives, or external memory banks, TaxoCon retains the observed coarse-class partition and uses an exponential moving average (EMA) teacher to redistribute positive-pair attraction according to similarity already encoded in the representation.
- We show that preserving rather than uniformly contracting within-class geometry leads to substantially stronger fine-grained representations under coarse supervision. Using a held-out-rank protocol in which finer-grained ranks are never used for representation learning, TaxoCon consistently preserves more local, linearly separable, and hierarchy-aligned structure across two hierarchical datasets and three backbone architectures, while maintaining competitive or improved performance at the supervised coarse rank.

2 Related Work

Pretrained Visual Representations. Pretrained visual models can encode rich semantic structure, including hierarchical relationships among visual categories Huang et al. (2021); Olson et al. (2025). However, fine-tuning can reshape the representation toward the downstream objective, potentially degrading semantic structure that is not directly supervised Kumar et al. (2022); Merlin et al. (2023). This is particularly relevant under coarse supervision, where distinctions among finer-grained categories are not explicitly represented in the training labels. This motivates our focus on preserving latent fine-grained structure during coarse-supervised fine-tuning, rather than assuming that such structure will survive adaptation.

Contrastive Representation Learning. Contrastive learning has become a powerful paradigm for learning discriminative and transferable visual representations Hadsell et al. (2006); van den Oord et al. (2018); He et al. (2020); Chen et al. (2020); Grill et al. (2020). Its central principle is to encourage semantically related samples to have similar representations while separating unrelated samples in the embedding space. This idea dates back to metric-learning approaches Hoffer & Ailon (2015); Schroff et al. (2015); Oh Song et al. (2016); Sohn (2016) based on contrastive objectives Hadsell et al. (2006) and has since been widely adopted in self-supervised visual representation

learning through methods such as InfoNCE van den Oord et al. (2018), MoCo He et al. (2020), SimCLR Chen et al. (2020), and BYOL Grill et al. (2020). Subsequent work has further demonstrated that the construction and weighting of positive and negative relationships can substantially affect the quality of the learned embedding space Chen et al. (2020); Yeh et al. (2022); Wang et al. (2019).

Supervised Contrastive Learning. Supervised Contrastive Learning (SupCon) Khosla et al. (2020) uses class labels to define positives, treating same-class samples as positives and others as negatives. While effective under flat-label supervision, coarse classes may contain visually and semantically distinct latent subgroups, so uniformly attracting all same-class samples can collapse fine-grained structure. Multi-Similarity Loss Wang et al. (2019) instead emphasizes informative pairs based on similarity and difficulty. TaxoCon differs by using stop-gradient similarities from a temporally smoothed EMA teacher to redistribute the positive learning signal within each coarse class.

Weighted Contrastive Losses. Several approaches move beyond binary positive treatment by weighting pairs based on similarity Wang et al. (2019); Manmatha et al. (2017). SCE Denize et al. (2023) uses a similarity distribution from a Siamese network to soften the hard positive–negative split of standard contrastive losses van den Oord et al. (2018); Chen et al. (2020); He et al. (2020), while X-CLR Sobal et al. (2024) replaces binary targets with continuous similarity-derived targets Wang et al. (2020); Robinson et al. (2021). These methods target self-supervised representation learning instead of latent fine-grained structure under coarse supervision. TaxoCon instead reweights same-class relationships during supervised contrastive fine-tuning using only coarse labels.

Coarse-to-Fine Representation Learning. Several methods directly target learning representations at a finer granularity than the labels available during training. Grafit Touvron et al. (2021) combines coarse-label supervision with an auxiliary instance-level self-supervised objective, while BuCSFR Shi et al. (2025) constructs a dendrogram from coarse-labelled data and explicitly learns separable fine-grained representations from the resulting hierarchy. These approaches are structurally different from TaxoCon, which introduces neither an auxiliary self-supervised objective nor an explicit clustering hierarchy. MaskCon Feng & Patras (2023) is more closely related: it also combines coarse labels with inter-sample similarities to soften uniform contrastive supervision. However, MaskCon constructs similarity-derived soft contrastive targets masked by coarse labels, whereas TaxoCon retains the supervised contrastive objective and uses stop-gradient similarities from a temporally smoothed EMA teacher to reweight same-label positives.

3 METHODS

3.1 SUPERVISED CONTRASTIVE LEARNING

Supervised Contrastive Learning (SupCon) Khosla et al. (2020) extends contrastive learning by using class labels to identify multiple positive examples for each anchor. For an anchor i, let $P(i)$ denote the set of samples sharing the same class label, and let $A(i)$ denote the set of samples considered in the contrastive denominator. For the SupCon baseline used in this work, each image contributes a single view, and $P(i)$ consists of the other mini-batch samples sharing the anchor's coarse training label. The SupCon objective is

$$\mathcal{L}_i^{\text{SupCon}} = -\frac{1}{|P(i)|} \sum_{p \in P(i)} \log \frac{\exp(z_i^\top z_p/\tau)}{\sum_{a \in A(i)} \exp(z_i^\top z_a/\tau)}. \tag{1}$$

where z_i and z_p are the normalized embeddings of the anchor and positive sample, respectively, and $\tau > 0$ is the contrastive temperature. The factor $1/|P(i)|$ assigns an equal weight to every positive pair, such that all samples sharing the same class label contribute equally to the loss. This uniform weighting treats every same-class pair as equally positive and encourages the model to reduce distances between samples that may be visually or semantically distinct at finer taxonomic levels. Under coarse taxonomic supervision, for example, instances belonging to the same *order* are all treated as equally positive even though they may belong to different *genera*. Consequently, the uniform positive objective can suppress fine-grained sub-structure within the coarse class.

3.2 SIMILARITY-WEIGHTED POSITIVES

Standard SupCon assigns the same weight to every positive pair within a coarse class. We instead redistribute the positive-pair learning signal according to similarity estimated by a slowly evolving teacher encoder. Specifically, for an anchor i, we define

$$\mathcal{L}_i = -\sum_{p \in P(i)} w_{ip} \log \frac{\exp(z_i^\top z_p/\tau)}{\sum_{a \in A(i)} \exp(z_i^\top z_a/\tau)}, \qquad \sum_{p \in P(i)} w_{ip} = 1, \tag{2}$$

where $P(i)$ denotes the set of positive samples for anchor i, and w_{ip} determines how strongly each positive pair contributes to the loss. The positive weights are obtained by applying a temperature-scaled softmax to teacher-estimated similarities:

$$w_{ip} = \frac{\exp(\tilde{s}_{ip}/\tau_p)}{\sum_{p' \in P(i)} \exp(\tilde{s}_{ip'}/\tau_p)}, \tag{3}$$

where $\tau_p > 0$ controls the sharpness of the positive weighting and distribution of the positive-pair learning signal. We consider two strategies for estimating $\tilde{s}_{ip}$. In the single-view variant, denoted *TaxoCon* (SV), the original image is encoded by the teacher (left in Eq. 4); in the multi-view variant, denoted *TaxoCon(MV)*, several independently augmented views of each image are passed through the teacher and their pairwise similarities are aggregated (right in Eq. 4):

$$\tilde{s}_{ip}^{\mathrm{SV}} = \tilde{z}_i^\top \tilde{z}_p, \qquad \tilde{s}_{ip}^{\mathrm{MV}} = \frac{1}{V^2} \sum_{v=1}^{V} \sum_{u=1}^{V} \left(\tilde{z}_i^{(v)}\right)^\top \tilde{z}_p^{(u)} \tag{4}$$

where $\tilde{z}_i$ and $\tilde{z}_p$ are normalized teacher embeddings, given V teacher views per image. The multi-view estimate aims to provide a more robust similarity estimate of the underlying relationship between positive samples. Consequently, same-class pairs estimated to be more similar receive larger weights, while less similar positive pairs receive smaller weights. This weighting does not remove the attraction between dissimilar positive pairs; rather, it redistributes optimization pressure among positive pairs allocating more positive-pair supervision to samples that are already close in representation space. This provides a mechanism for reinforcing local structure within coarse classes while retaining the inter-class separation induced by contrastive learning. As $\tau_p \to 0$, the weights increasingly concentrate on the most similar positive pairs. As $\tau_p \to \infty$, the weights become uniform and Eq. 2 reduces exactly to the standard SupCon objective. Thus, SupCon is recovered as a limiting case of our formulation.

3.3 EMA TEACHER FOR STABLE SIMILARITY ESTIMATES

We compute the positive weights using a separate EMA teacher encoder, as shown in Eq. 3, rather than the online encoder. For both the single-view and multi-view variants, the teacher is used only to estimate the similarities that determine the positive weights; the optimization objective for the online representation remains Eq. 2. The teacher parameters are updated from the online encoder using an exponential moving average:

$$\theta_{\mathrm{EMA}} \leftarrow m\theta_{\mathrm{EMA}} + (1-m)\theta_{\mathrm{online}}, \tag{5}$$

where $m \in [0, 1)$ is the EMA momentum.

The teacher representations and positive weights are stop-gradient, providing a temporally smoothed estimate of similarity without receiving gradients from the contrastive objective. This separation avoids feedback loops in which transient increases in online similarity amplify pair weights and further reinforce those similarities, while also reducing instability from rapidly changing online representations early in training.

The multi-view variant further averages similarity across stochastic augmentations, reducing dependence on any single view. EMA therefore provides temporal smoothing across optimization steps, while multi-view aggregation provides augmentation-level smoothing within each step. Neither variant uses fine-grained labels; the teacher supplies only representation-derived similarity estimates and receives no fine-grained supervision, allowing TaxoCon to preserve within-class structure without additional fine-grained annotations.

3.4 Relation to Existing Coarse-to-Fine Methods

MaskCon Feng & Patras (2023) and TaxoCon both exploit inter-sample similarity under coarse supervision, but they incorporate it differently. MaskCon uses similarities from the current representation to construct a coarse-label-masked soft target distribution and combines this objective with a self-supervised contrastive loss over augmented views. TaxoCon instead retains the standard supervised contrastive objective and uses stop-gradient similarities from an EMA teacher only to redistribute attraction among same-label positives. In TaxoCon(MV), augmentation is used to obtain a more steady and robust similarity estimate rather than to define an additional self-supervised objective. TaxoCon further simplifies the framework by operating entirely within the current mini-batch, without requiring a memory bank or queue. Moreover, as the positive-weighting temperature $\tau_p \to \infty$, the similarity weights become uniform and TaxoCon reduces exactly to standard SupCon making TaxoCon a direct generalization of SupCon. Grafit Touvron et al. (2021) and BuCSFR Shi et al. (2025) address the same coarse-to-fine setting through more structurally different mechanisms. Grafit combines coarse-label supervision with an auxiliary instance-level self-supervised objective, while BuCSFR explicitly constructs a dendrogram from coarse-labelled data; TaxoCon instead modifies the supervised contrastive objective directly, requiring neither an auxiliary self-supervised loss nor an explicitly constructed hierarchy.

4 Experiments

4.1 Dataset

We conduct experiments on two hierarchically labelled fine-grained recognition datasets: *iNaturalist 2021* Van Horn et al. (2021) and *FGVC-Aircraft* Maji et al. (2013). The two datasets provide complementary testbeds for evaluating whether coarse-label supervision can preserve meaningful fine-grained structure across different domains and hierarchy types. The iNaturalist 2021 labels follow the complete Linnaean biological taxonomic hierarchy: *kingdom* $\succ$ *phylum* $\succ$ *class* $\succ$ *order* $\succ$ *family* $\succ$ *genus* $\succ$ *species*. FGVC-Aircraft provides a complementary non-biological hierarchy, with aircraft organised as *manufacturer* $\succ$ *family* $\succ$ *variant*. This offers a complementary setting in which fine-grained distinctions are defined by subtle visual differences within a human-designed categorical hierarchy rather than by biological taxonomy. Evaluating on both datasets therefore allows us to test whether the proposed approach generalises across substantially different domains, hierarchy depths, and sources of fine-grained variation. Dataset details are provided in Appendix C.

4.2 Held-out rank protocol

For iNaturalist 2021, we use *order* as the supervised training rank, denoted by r_{train}, and evaluate the learned embeddings at the finer-grained ranks $r_{\text{test}} \in \{\textit{family}, \textit{genus}, \textit{species}\}$. For FGVC-Aircraft, we use *manufacturer* as r_{train} and evaluate at $r_{\text{test}} \in \{\textit{family}, \textit{variant}\}$. In both datasets, the pretrained backbones are fine-tuned using only r_{train} ranks. r_{test} are never seen in representation fine-tuning and hyperparameter selection and are only used afterwards to see how much fine-grained structure is retained. For linear-probing, finer-grained training ranks are used to fit the probe, while evaluation is performed on the held-out test split. This protocol tests whether a method can preserve fine-grained hierarchical structure without sacrificing performance at the supervised rank. Therefore, any preservation or improvement at r_{test} cannot be attributed to direct fine-grained supervision during representation learning.

4.3 Experimental Settings

Pretrained backbones & Baselines. To assess whether the observed effects are independent of backbone architecture, we evaluate three pretrained models: *ResNet-50* He et al. (2016), *ViT-S/16* Dosovitskiy et al. (2021), and *Swin-T* Liu et al. (2021). Details are provided in Appendix D.

We compare TaxoCon against several complementary baselines. We include standard supervised objectives: Supervised Contrastive Learning (***SupCon***), to which TaxoCon reduces as $\tau_p \to \infty$; Multi-Similarity loss (***MS-loss***), a metric-learning objective that jointly considers multiple positive and negative pairs; and ***Cross Entropy***, a standard classification objective. Among existing coarse-to-fine methods, we include ***MaskCon*** as the closest methodological baseline. Grafit and BuCSFR

address the same broader problem but use substantially different learning mechanisms; our comparison therefore focuses on methods that more directly isolate the effect of similarity-aware contrastive learning. Finally, we report the pretrained model without fine-tuning as a ***Frozen Backbone***, providing a reference for how each fine-tuning strategy affects pre-existing fine-grained structure.

TaxoCon variants. We evaluate three variants of TaxoCon. ***TaxoCon(SV)*** uses the original image to compute similarity weights (Eq. 4); ***TaxoCon(MV)*** uses two independently augmented views (Eq. 4). We additionally evaluate ***TaxoCon w/o EMA***, an ablation that removes the EMA teacher and computes single-view similarity weights directly from the online embeddings.

Evaluation. We conduct evaluation at three complementary levels. First, we assess local neighborhood structure using *Recall@1* and *Recall@5*. Recall@K measures the proportion of query samples for which at least one sample sharing the same evaluation label occurs among the K nearest neighbors in the embedding space. Second, we measure linear separability using linear-probing. Lastly, we evaluate global hierarchical structure using two complementary measures: *Spearman* correlation, which quantifies how well pairwise embedding distances preserve the ordering of taxonomic distances Saraçli et al. (2013); D'haeseleer (2005), and *Dendrogram–Taxonomy Agreement*, which measures how closely hierarchical clustering of the learned embeddings recovers the known taxonomic hierarchy. Details are provided in Appendix B.

Hyperparameter. Main results reported in Section 5 use $\tau_p = 0.1$ and $m = 0.95$ for TaxoCon(SV) / TaxoCon(MV); $\tau_p = 0.3$ for TaxoCon w/o EMA. These hyperparameters are selected exclusively on the iNaturalist 2021 dataset using Recall@1 at the supervised training rank (*order*) on a validation split comprising 10% of the training data. This avoids using finer-grained labels for model selection and therefore prevents information leakage from the evaluation hierarchy. Labels at the finer-grained ranks are used only for post-hoc evaluation. After selection, each model is retrained on the full iNaturalist 2021 training set using the selected configuration. The resulting hyperparameter configurations are treated as the default settings and transferred directly to FGVC-Aircraft without dataset-specific tuning.

4.4 iNaturalist Experimental Splits

For iNaturalist 2021, we evaluate both *per-group* and *full-hierarchy* training settings.

In *per-group* setting, models are trained separately within each of nine class groups (e.g., *Birds*, *Insects*, etc.) (see Appendix C). We think this separation is important because the dataset is highly imbalanced across these groups: larger groups contain substantially more images and categories and could otherwise dominate performance. Restricting each experiment to one group reduces this source of class-size and hierarchy imbalance and keeps the learning and evaluation problem more locally comparable. It also allows performance to be examined separately across biologically distinct groups rather than being dominated by the largest groups. We report results for each group individually in Appendix F and macro-average results across all nine groups in the main report.

In *full-hierarchy* setting, We fine-tune a backbone over the complete iNaturalist 2021 hierarchy to test whether the relative advantage of TaxoCon over MaskCon persists without group-specific training. Due to computational constraints, this full-hierarchy comparison is performed only with ResNet-50 at 32×32 resolution on TaxoCon(MV) and MaskCon and is treated as a robustness experiment rather than a directly comparable benchmark to the 224×224 main results. Due to computational constraints, MaskCon is evaluated in both settings using ResNet-50 only. For FGVC-Aircraft, we use the complete dataset hierarchy.

5 Results

5.1 Results for iNaturalist 2021

Macro-averaged Results Table 1 shows that on average, TaxoCon consistently improves fine-grained structure preservation under coarse supervision, with TaxoCon(MV) generally performing best. The gains are strongest for the transformer backbones and at the finer-grained ranks. For ViT-S/16, TaxoCon(MV) improves over the strongest non-TaxoCon baseline by $+9.00\%$, $+8.64\%$,

Table 1: macro-averaged Recall@1/5, linear-probing (Lp), Spearman correlation (Sp.), and Dendrogram–Taxonomy Agreement (Dr.) across backbones and methods for the iNaturalist 2021 dataset at different ranks.

		Order(r_{train})			Family			Genus			Species				
Backbone	**Method**	Recall@1	Recall@5	LP	Recall@1	Recall@5	LP	Recall@1	Recall@5	LP	Recall@1	Recall@5	LP	Sp.	Dr.
ResNet-50	Frozen Backbone	64.98	83.02	69.32	29.42	55.96	43.74	13.86	32.35	21.74	8.01	20.58	12.22	17.73	**18.02**
	SupCon	69.38	83.16	70.72	28.87	56.05	38.32	12.29	30.54	19.64	7.23	19.12	10.37	18.58	17.07
	MS-loss	68.15	80.46	68.37	28.86	55.13	37.46	12.67	30.23	20.66	7.17	18.76	9.88	16.32	14.88
	Cross Entropy	64.66	79.60	66.92	19.98	44.88	27.72	6.57	19.24	10.44	3.17	10.31	4.52	14.91	14.04
	MaskCon	76.04	88.47	76.21	38.24	66.02	45.09	20.88	42.56	27.50	13.35	29.87	16.94	20.18	17.29
	TaxoCon w/o EMA	71.03	83.09	71.98	29.71	56.77	38.42	12.63	31.43	19.91	7.49	19.65	10.12	18.84	17.29
	TaxoCon(SV)	72.19	84.32	73.17	33.70	60.38	40.89	15.82	35.95	22.40	9.37	23.50	12.93	19.57	17.55
	TaxoCon(MV)	**78.86**	**88.92**	**78.47**	**41.65**	**68.53**	**48.09**	**22.42**	**45.97**	**28.91**	**14.18**	**31.76**	**17.15**	**21.60**	**19.38**
	$\Delta_{\textit{TaxoCon gain}}{}^{*}$	**+2.82**	**+0.45**	**+2.26**	**+3.41**	**+2.51**	**+3.00**	**+1.54**	**+3.41**	**+1.41**	**+0.83**	**+1.89**	**+0.21**	**+1.42**	**+1.36**
ViT-S/16	Frozen Backbone	71.56	88.66	72.09	45.10	69.54	50.43	26.99	50.20	37.83	18.09	37.19	21.72	22.90	20.17
	SupCon	78.63	89.38	77.81	40.00	66.26	45.52	21.16	42.87	28.27	12.98	29.24	15.47	19.39	18.40
	MS-loss	80.22	89.04	77.86	47.83	71.10	49.12	28.87	50.90	33.38	18.71	37.34	20.96	20.04	17.63
	Cross Entropy	66.92	83.09	68.27	24.06	50.86	31.13	9.43	24.96	14.58	4.81	14.09	6.53	15.06	14.06
	TaxoCon w/o EMA	79.55	89.09	78.49	40.85	66.50	46.35	21.59	42.70	29.13	13.02	29.12	16.47	18.63	17.46
	TaxoCon(SV)	81.89	91.35	81.60	52.32	76.02	55.57	33.66	57.58	40.02	23.02	44.54	26.41	23.27	21.64
	TaxoCon(MV)	**84.99**	**91.97**	**83.84**	**56.83**	**78.33**	**57.98**	**37.51**	**61.06**	**42.30**	**26.31**	**47.79**	**28.45**	**24.53**	**22.66**
	$\Delta_{\textit{TaxoCon gain}}$	**+4.77**	**+2.59**	**+5.98**	**+9.00**	**+7.23**	**+7.55**	**+8.64**	**+10.16**	**+4.47**	**+7.60**	**+10.45**	**+6.73**	**+1.63**	**+2.49**
Swin-T	Frozen Backbone	69.80	86.28	71.77	39.51	65.75	48.27	23.69	46.92	34.52	13.95	30.43	19.95	23.14	21.58
	SupCon	80.76	90.64	81.11	39.11	67.49	45.80	19.91	42.29	27.10	12.48	29.19	14.50	21.73	20.75
	MS-loss	82.45	90.00	80.88	47.94	72.16	51.32	28.16	51.14	33.14	18.30	37.20	20.39	21.33	19.66
	Cross Entropy	75.52	87.48	77.71	34.19	60.70	44.93	15.32	35.33	22.99	8.18	21.90	13.20	18.98	17.43
	TaxoCon w/o EMA	82.26	90.24	80.93	43.48	68.77	47.13	23.50	45.14	28.72	14.78	31.54	15.58	17.51	16.09
	TaxoCon(SV)	84.15	91.69	83.31	52.27	75.72	54.68	31.88	55.96	36.12	21.16	42.23	22.57	23.44	22.21
	TaxoCon(MV)	**87.30**	**92.39**	**86.31**	**55.84**	**77.56**	**59.19**	**34.97**	**58.94**	**40.98**	**23.68**	**45.16**	**27.17**	**24.64**	**23.10**
	$\Delta_{\textit{TaxoCon gain}}$	**+4.85**	**+1.75**	**+5.20**	**+7.90**	**+5.40**	**+7.87**	**+6.81**	**+7.80**	**+6.46**	**+5.38**	**+7.96**	**+6.78**	**+1.50**	**+1.52**

$*$ This $\Delta_{\textit{TaxoCon gain}}$ includes comparison with MaskCon.

and $+7.60\%$ in Recall@1 at finer ranks, with similar gains at Recall@5. Swin-T shows a similar pattern with comparable gains. The increasing advantage at finer ranks suggests that TaxoCon is not merely improving coarse discrimination, but better retaining local structure within coarse classes that remains useful for distinguishing unseen finer-grained categories. TaxoCon(MV) also achieves the highest Spearman and Dendrogram–Taxonomy Agreement scores among the methods evaluated for each backbone, indicating that these local improvements are accompanied by better global alignment between the learned geometry and the ground-truth taxonomy. For ResNet-50, MaskCon is the strongest competing baseline, yet TaxoCon(MV) exceeds it across all reported metrics. TaxoCon(SV) generally provides intermediate gains, while removing the EMA teacher substantially weakens performance. Together, this pattern suggests that stable teacher-based similarity estimates preserve informative within-class relationships, while multi-view aggregation further reduces sensitivity to augmentation-specific variation.

Per-group Results Figure 3 shows that these aggregate gains extend across individual groups. TaxoCon(MV) achieves the strongest *Species*-level Recall@1 for most groups, with clear improvements for seven groups. The gain is especially pronounced for *Plants*, where baseline fine-grained retrieval is low. *Mollusks* and *Fungi* provide exceptions, with MaskCon slightly outperforming TaxoCon(MV), showing that the advantage is not uniform across taxonomic branches. Overall, the per-group results indicate that the macro-level improvement arises across diverse biological groups rather than being driven by a single group. Full *per-group* results can be found in Appendix F.

Full-hierarchy Evaluation. To determine whether the observed advantage depends on training separate group-specific models, we additionally compare TaxoCon(MV) and MaskCon using ResNet-50 trained over the complete iNaturalist 2021 hierarchy at 32×32 resolution. As shown in Table 2, TaxoCon outperforms MaskCon across all metrics. This may reflect TaxoCon's use of EMA-smoothed similarity estimates to reweight only same-label positives, preserving local within-class structure without imposing MaskCon's similarity-derived soft targets over the full contrastive distribution. At the supervised *Order* rank, TaxoCon improves Recall@1 from 19.95% to 23.76% and linear-probing from 26.93% to 29.71%. The advantage persists at the finer-grained ranks. Spearman taxonomic-distance correlation and Dendrogram–Taxonomy Agreement also increase

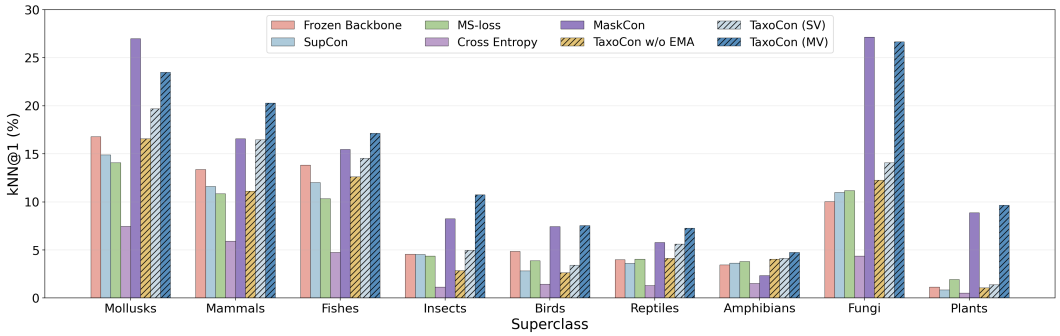

Figure 3: *per-group* Recall@1 at *Species* rank for iNaturalist 2021 on ResNet-50.

Table 2: low-resolution *full-hierarchy* iNaturalist 2021 comparison between MaskCon and TaxoCon(MV) using ResNet-50 at 32×32 resolution, evaluating relative robustness under full-hierarchy training.

	Order (r_{train})			**Family**			**Genus**			**Species**				
Method	Recall@1	Recall@5	Lp	Recall@1	Recall@5	Lp	Recall@1	Recall@5	Lp	Recall@1	Recall@5	Lp	Sp.	Dr.
MaskCon	19.95	42.23	26.93	7.38	21.77	9.86	1.98	6.61	2.51	0.99	3.47	0.77	7.34	7.17
TaxoCon(MV)	**23.76**	**45.48**	**29.71**	**9.40**	**25.29**	**11.59**	**2.70**	**8.68**	**3.04**	**1.27**	**4.34**	**0.96**	**8.32**	**7.23**
$\Delta_{\textit{TaxoCon gain}}$	**+3.81**	**+3.25**	**+2.78**	**+2.02**	**+3.52**	**+1.73**	**+0.72**	**+2.07**	**+0.53**	**+0.28**	**+0.87**	**+0.19**	**+0.98**	**+0.06**

marginally. Because the 32×32 input resolution substantially reduces fine-grained visual detail, the consistently positive relative differences provide evidence that TaxoCon's advantage over MaskCon persists when training a single model across the full hierarchy.

5.2 Results on FGVC-Aircraft

Table 3 shows that TaxoCon generalizes effectively to FGVC-Aircraft, with its strongest advantages appearing at *Family* and *Variant*. Relative to the strongest non-TaxoCon baseline for each metric, TaxoCon achieves positive gains in nearly all cases across all backbones. At the finer-grained ranks, the improvements in Recall@1 range from $+0.91\%$ to $+4.63\%$ at *Family* and from $+1.11\%$ to $+3.96\%$ at *Variant*. Similar gains are observed for Recall@5 and linear-probing, indicating that TaxoCon improves both local neighborhood structure and linear separability of the latent fine-grained categories. TaxoCon also improves the global organization of the embedding space, achieving higher Spearman correlation and Dendrogram–Taxonomy Agreement than the strongest non-TaxoCon baselines for all three backbones. This suggests that the learned representation better reflects the underlying hierarchy rather than merely improving discrimination at individual ranks.

Across the different TaxoCon formulations, no single variant dominates every metric. TaxoCon(MV) is generally strongest on local neighborhood retrieval and linear-probing performance, whereas TaxoCon(SV) and TaxoCon w/o EMA occasionally achieve better hierarchy-alignment scores or supervised-rank performance. This variation suggests that the different formulations emphasize complementary properties of the embedding space. Nevertheless, the consistent positive $\Delta_{\textit{TaxoCon gain}}$ show that the broader TaxoCon framework remains competitive with or improves upon strong alternatives such as MaskCon and MS-loss. Overall, these results reinforce the findings on iNaturalist 2021. The advantages of TaxoCon are most evident at the finer-grained ranks, where it consistently improves all metrics across substantially different backbone architectures.

5.3 Ablation studies

We ablate $\tau_p \in \{0.05, 0.1, 0.3, 0.5, 0.7, 1.0\}$ and $m \in \{0.9, 0.95, 0.999\}$ on *iNaturalist 2021*. This allows us to isolate the effect of each hyperparameter on the preservation of fine-grained taxonomic structure and on performance at the supervised rank. Performance is also relatively stable across $m \in 0.9, 0.95, 0.999$, with $m = 0.95$ providing strong or near-best performance across most back-

Table 3: Recall@1/5, linear-probing (Lp), Spearman correlation (Sp.), and Dendrogram–Taxonomy Agreement (Dr.) on FGVC-Aircraft across backbones and methods at the supervised *Manufacturer* rank and finer-grained *Family* and *Variant* ranks. $\Delta_{\textit{TaxoCon gain}}$ denotes the difference between the best-performing TaxoCon variant and the best non-TaxoCon method for each metric.

		Manufacturer(r_{train})			**Family**			**Variant**				
Backbone	**Method**	Recall@1	Recall@5	Lp	Recall@1	Recall@5	Lp	Recall@1	Recall@5	Lp	Sp.	Dr.
ResNet-50	Frozen Backbone	42.84	73.06	51.27	30.48	55.57	41.68	23.13	42.69	31.48	18.10	15.58
	SupCon	60.64	81.79	65.52	40.62	66.40	52.32	29.67	54.61	36.73	19.22	17.18
	MS-loss	57.13	79.18	61.47	39.33	65.59	50.67	28.68	53.41	36.13	16.83	15.70
	Cross Entropy	41.46	72.67	52.92	24.99	50.98	37.93	17.70	38.10	28.49	15.98	14.60
	MaskCon	57.22	84.91	65.07	43.89	70.00	56.82	34.38	58.30	43.18	17.67	15.95
	TaxoCon w/o EMA	66.64	84.04	**72.26**	45.63	70.81	58.47	33.27	59.59	43.03	20.08	**19.08**
	TaxoCon(SV)	**66.88**	83.74	71.21	47.52	72.76	57.57	35.10	60.97	42.28	**20.27**	18.92
	TaxoCon(MV)	62.59	**87.76**	67.02	**48.52**	**75.25**	**59.07**	**38.34**	**64.06**	**45.58**	19.66	18.77
	$\Delta_{\textit{TaxoCon gain}}$	**+6.24**	**+2.85**	**+6.74**	**+4.63**	**+5.25**	**+2.25**	**+3.96**	**+5.76**	**+2.40**	**+1.05**	**+1.90**
ViT-S/16	Frozen Backbone	44.07	76.96	49.18	32.82	59.32	44.53	25.62	46.95	32.98	18.43	16.05
	SupCon	73.54	88.87	73.91	49.69	76.15	57.12	36.51	63.61	43.48	21.03	20.51
	MS-loss	79.45	90.25	79.91	55.09	79.12	63.56	41.46	67.96	51.01	21.38	19.51
	Cross Entropy	64.99	85.03	63.87	39.48	67.60	48.13	28.44	55.33	35.38	19.97	19.61
	MaskCon	84.76	**93.79**	83.36	64.95	85.24	71.21	47.33	73.01	53.37	22.50	21.70
	TaxoCon w/o EMA	78.22	90.07	77.21	54.25	78.25	63.87	41.70	66.01	49.03	21.20	21.32
	TaxoCon(SV)	79.99	91.87	77.81	61.18	82.78	64.62	48.06	72.70	51.42	21.67	21.11
	TaxoCon(MV)	**86.92**	93.67	**86.96**	**65.86**	**85.27**	**74.06**	**48.72**	**74.92**	**55.77**	**22.83**	**22.35**
	$\Delta_{\textit{TaxoCon gain}}$	**+2.16**	-0.12	**+3.60**	**+0.91**	**+0.03**	**+2.85**	**+1.39**	**+1.91**	**+2.40**	**+0.33**	**+0.65**
Swin-T	Frozen Backbone	41.10	74.23	58.02	29.34	57.01	46.48	22.50	44.22	37.63	18.06	17.20
	SupCon	88.30	94.63	87.86	56.14	82.54	69.42	44.37	70.03	55.02	22.85	23.35
	MS-loss	90.52	**95.23**	90.25	66.01	86.83	76.76	52.42	76.87	60.57	23.19	22.83
	Cross Entropy	85.93	92.68	85.61	57.25	81.46	68.07	42.36	70.03	51.87	22.32	22.05
	MaskCon	88.12	94.51	86.21	68.47	86.56	79.01	52.06	75.46	59.07	22.89	22.59
	TaxoCon w/o EMA	90.31	94.60	88.01	67.69	86.47	76.46	52.66	**77.23**	60.47	22.87	23.64
	TaxoCon(SV)	**91.66**	**95.23**	**91.75**	69.01	87.49	79.01	**53.53**	75.22	60.87	**23.37**	**24.03**
	TaxoCon(MV)	90.64	94.96	91.00	**69.43**	**87.70**	**80.06**	51.97	76.57	**61.17**	23.21	22.97
	$\Delta_{\textit{TaxoCon gain}}$	**+1.14**	0.00	**+1.50**	**+0.96**	**+0.87**	**+1.05**	**+1.11**	**+0.36**	**+0.60**	**+0.18**	**+0.68**

bones and metrics. Overall, the gains remain consistent with the configuration selected using supervised *order*-level validation. The complete results are reported in Appendices E.1 and E.2.

6 Conclusion

We introduced **TaxoCon**, an EMA-guided similarity-weighted supervised contrastive objective for preserving fine-grained structure during coarse-supervised fine-tuning. We frame coarse-supervised fine-tuning as a representation-preservation problem: preserving fine-grained structure already encoded in pretrained representations. TaxoCon addresses this problem by retaining the observed coarse-class partition while redistributing positive-pair attraction according to within-class similarity estimated by an EMA teacher. Similar positives are therefore attracted more strongly and dissimilar positives less strongly, reducing the uniform contraction of heterogeneous samples within a coarse class without requiring fine-grained labels, an auxiliary self-supervised objective, or an external memory bank.

TaxoCon consistently preserves stronger fine-grained structure than the evaluated baselines on three pretrained backbone architectures and two datasets while maintaining competitive or improved performance at the supervised coarse rank. These results demonstrate that useful fine-grained structure already present in pretrained representations can be exploited as a signal for preserving within-class geometry during coarse-supervised adaptation. The effectiveness of this strategy nevertheless depends on the quality of the latent structure available in the pretrained representation. Future work should therefore investigate reliability-aware similarity estimation and adaptive weighting strategies that can determine when and how strongly existing representation structure should be preserved.

Overall, our results support a shift in perspective for coarse-to-fine representation learning: fine-grained structure need not always be rediscovered or imposed from coarse supervision. When such structure is already encoded by pretraining, preserving it during adaptation can provide a direct path to stronger fine-grained representations.

AI USE DISCLOSURE

Generative AI tools were used to assist in identifying potentially relevant literature and to improve the clarity and readability of the manuscript. They were not used to develop the research methodology, formulate hypotheses or mathematical claims, implement the proposed method, design experiments, analyse data, or interpret experimental results. All identified literature was independently checked by the authors, and all AI-assisted edits were reviewed before inclusion. The authors take full responsibility for the final content of the manuscript.

REPRODUCIBILITY STATEMENT

We have taken several steps to support reproducibility. The proposed method is fully specified in Section 3, with additional implementation and numerical details provided in Appendix A. The experimental protocol and evaluation metrics are described in Section 4 and Appendix B, while the dataset and training configurations are documented in Appendices C and D. Complete hyperparameter ablations and per-group results are provided in Appendices E.1, E.2, and F, respectively. Together, these sections provide the methodological, implementation, training, and evaluation information needed to reproduce the reported experiments.

REFERENCES

Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *Proceedings of the 37th International Conference on Machine Learning (ICML)*, volume 119, pp. 1597–1607, 2020.

Yin Cui, Yang Song, Chen Sun, Andrew Howard, and Serge Belongie. Large scale fine-grained categorization and domain-specific transfer learning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 4109–4118, 2018.

Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 248–255, 2009.

Julien Denize, Jaonary Rabarisoa, Astrid Orcesi, Romain Hérault, and Stéphane Canu. Similarity contrastive estimation for self-supervised soft contrastive learning. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pp. 2706–2716, 2023.

Patrik D'haeseleer. How does gene expression clustering work? *Nature Biotechnology*, 23(12): 1499–1501, 2005. doi: 10.1038/nbt1205-1499.

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. URL https://openreview.net/forum?id=YicbFdNTTy.

Chen Feng and Ioannis Patras. Maskcon: Masked contrastive learning for coarse-labelled dataset. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 19913–19922, 2023. doi: 10.1109/CVPR52729.2023.01907.

Jean-Bastien Grill, Florian Strub, Florent Altché, Corentin Tallec, Pierre Richemond, Elena Buchatskaya, Carl Doersch, Bernardo Avila Pires, Zhaohan Guo, Mohammad Gheshlaghi Azar, et al. Bootstrap your own latent: A new approach to self-supervised learning. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 33, pp. 21271–21284, 2020.

Raia Hadsell, Sumit Chopra, and Yann LeCun. Dimensionality reduction by learning an invariant mapping. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 1735–1742, 2006. doi: 10.1109/CVPR.2006.100.

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*, pp. 770–778, 2016. doi: 10.1109/CVPR.2016.90.

Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. Momentum contrast for unsupervised visual representation learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 9729–9738, 2020.

Elad Hoffer and Nir Ailon. Deep metric learning using triplet network. *Similarity-Based Pattern Recognition*, pp. 84–92, 2015.

Taicheng H. Huang, Zonglei Zhen, and Jia Liu. Semantic relatedness emerges in deep convolutional neural networks designed for object recognition. *Frontiers in Computational Neuroscience*, 15, 2021.

Prannay Khosla, Piotr Teterwak, Chen Wang, Aaron Sarna, Yonglong Tian, Phillip Isola, Aaron Maschinot, Ce Liu, and Dilip Krishnan. Supervised contrastive learning. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 33, pp. 18661–18673, 2020.

Ananya Kumar, Aditi Raghunathan, Robbie Matthew Jones, Tengyu Ma, and Percy Liang. Fine-tuning can distort pretrained features and underperform out-of-distribution. In *International Conference on Learning Representations*, 2022.

Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 10012–10022, 2021. doi: 10.1109/ICCV48922.2021.00986.

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

R. Manmatha, Feng Hua, and R. Chellappa. Sampling matters in deep embedding learning. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pp. 2840–2848, 2017.

Gabriele Merlin, Vedant Nanda, Ruchit Rawal, and Mariya Toneva. What happens during finetuning of vision transformers: An invariance based investigation. In *Proceedings of the 2nd Conference on Lifelong Learning Agents*, volume 232, pp. 601–619. PMLR, 2023.

Hyun Oh Song, Yu Xiang, Stefanie Jegelka, and Silvio Savarese. Deep metric learning via lifted structured feature embedding. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 4004–4012, 2016.

Matthew Lyle Olson, Musashi Hinck, Neale Ratzlaff, Changbai Li, Phillip Howard, Vasudev Lal, and Shao-Yen Tseng. Analyzing hierarchical structure in vision models with sparse autoencoders. In *CVPR Workshops*, pp. 4874–4878, 2025.

Tal Ridnik, Emanuel Ben-Baruch, Asaf Noy, and Lihi Zelnik-Manor. Imagenet-21k pretraining for the masses. *arXiv preprint arXiv:2104.10972*, 2021.

Joshua Robinson, Ching-Yao Chuang, Suvrit Sra, and Stefanie Jegelka. Contrastive learning with hard negative samples. In *International Conference on Learning Representations*, 2021.

Sinan Saraçli, Nurhan Doğan, and İsmet Doğan. Comparison of hierarchical cluster analysis methods by cophenetic correlation. *Journal of Inequalities and Applications*, 2013:203, 2013. doi: 10.1186/1029-242X-2013-203.

Florian Schroff, Dmitry Kalenichenko, and James Philbin. Facenet: A unified embedding for face recognition and clustering. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 815–823, 2015.

Guanghui Shi, Xuefeng Liang, Wenjie Li, and Xiaoyu Lin. Learning separable fine-grained representation via dendrogram construction from coarse labels for fine-grained visual recognition. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 870–879, 2025.

Vlad Sobal, Mark Ibrahim, Randall Balestriero, Vivien Cabannes, Diane Bouchacourt, Pietro Astolfi, Kyunghyun Cho, and Yann LeCun. X-sample contrastive loss: Improving contrastive learning with sample similarity graphs. *arXiv preprint arXiv:2407.18134*, 2024.

Kihyuk Sohn. Improved deep metric learning with multi-class N-pair loss objective. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 29, 2016.

Hugo Touvron, Alexandre Sablayrolles, Matthijs Douze, Matthieu Cord, and Hervé Jégou. Grafit: Learning fine-grained image representations with coarse labels. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 874–884, 2021.

Aäron van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*, 2018.

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*, pp. 12884–12893, 2021. doi: 10.1109/CVPR46437.2021.01269.

Xin Wang et al. Cross-view training for semi-supervised learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020.

Xun Wang, Xintong Han, Weilin Huang, Dengke Dong, and Matthew R. Scott. Multi-similarity loss with general pair weighting for deep metric learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 5022–5030, 2019. doi: 10.1109/CVPR.2019.00516.

Chun-Hsiao Yeh, Cheng-Yao Hong, Yen-Chang Hsu, Tyng-Luh Liu, and Yu-Chiang Frank Wang. Decoupled contrastive learning. In *European Conference on Computer Vision*, pp. 668–684. Springer, 2022.

A Implementation and aggregation

Several details are required for a numerically sound implementation. First, both the contrastive logits $z_i^\top z_a/\tau$ and the weighting logits s_{ip}/τ_p are stabilized by per-row maximum subtraction prior to exponentiation, which leaves both softmaxes invariant while bounding the exponentiated arguments. Second, entries excluded by the mask in equation 3—self-comparisons and all negatives—are set to a large negative constant before the softmax, so that they receive numerically zero weight; the self-comparison is likewise excluded from the denominator of equation 2. Third, a small constant ϵ is added inside logarithms to guard against underflow. Finally, an anchor contributes to the batch aggregate only if it possesses at least one positive, since equation 3 is undefined when $P(i) = \emptyset$. The batch loss is therefore the mean over the surviving anchors,

$$\mathcal{L} = \frac{1}{|\mathcal{I}|} \sum_{i \in \mathcal{I}} \mathcal{L}_i, \qquad \mathcal{I} = \big\{\, i \in \{1, \dots, N\} : |P(i)| > 0 \,\big\}, \tag{6}$$

which leaves the loss well defined for batches containing singleton classes and renders the scale of $\mathcal{L}$ independent of how many such singletons a given batch happens to contain. Apart from the weighting in equation 3, TaxoCon introduces no additional parameters, no additional forward passes beyond the EMA teacher, and no change to the asymptotic $\mathcal{O}(N^2 d)$ cost of the pairwise similarity computation.

B Taxonomic Structure Metrics

The metrics in this section compare the geometry of the learned embedding space with the known ground-truth hierarchy. Given the same set of n encoded test images, we compare pairwise distances induced by the learned representation with pairwise distances defined by the corresponding dataset hierarchy. Intuitively, these metrics ask: *if two organisms are taxonomically more distant, are they also represented as more distant in the embedding space?*

B.1 NOTATION

Let $\{x_i\}_{i=1}^n$ denote the evaluated test images and let $z_i = f_\theta(x_i) \in \mathbb{R}^d$ be the ℓ_2-normalized embedding produced by the evaluated encoder f_θ.

Embedding distances. We use cosine distance:

$$d_{ij}^{\text{emb}} = 1 - \frac{\langle z_i, z_j \rangle}{\|z_i\| \, \|z_j\|}, \qquad 1 \le i < j \le n. \tag{7}$$

Taxonomic distances. Each image carries a hierarchical path $t_i = (t_i^{(1)}, \ldots, t_i^{(L)})$ over the L evaluated ranks. For iNaturalist 2021 these are *Order, Family, Genus*, and *Species*; for FGVC-Aircraft they are *Manufacturer, Family*, and *Variant*. Writing $t_i^{(1:r)}$ for the cumulative path up to rank r (so that identical names under different ancestors do not collapse), the depth of the lowest common ancestor (LCA) of i and j is

$$\text{lca}(i,j) = \sum_{r=1}^{L} \mathbf{1}\left[t_i^{(1:r)} = t_j^{(1:r)}\right], \tag{8}$$

and the taxonomic distance is

$$d_{ij}^{\text{tax}} = L - \text{lca}(i,j). \tag{9}$$

This defines an ultrametric over the evaluated taxonomy. For example, two images from the same *species* have distance 0, two *species* within the same genus have distance 1, and two images sharing only the coarsest evaluated rank have distance $L - 1$.

The pairwise embedding and taxonomic distances are flattened into condensed vectors

$$\mathbf{u} = (d_{ij}^{\text{emb}})_{i<j}, \qquad \mathbf{v} = (d_{ij}^{\text{tax}})_{i<j},$$

each of length $N = \binom{n}{2}$.

B.2 SPEARMAN TAXONOMIC DISTANCE CORRELATION

The Spearman coefficient measures the monotonic association between embedding distance and taxonomic distance. It is computed as the Pearson correlation between their ranks:

$$\rho_S; =; \text{corr}\left(\text{rank}(\mathbf{u}), \text{rank}(\mathbf{v})\right). \tag{10}$$

Tied values receive their average rank. Since $\mathbf{v}$ takes only a small number of discrete values corresponding to taxonomic separation, ties are common and are handled accordingly. Spearman correlation is invariant to monotonic transformations of the embedding distances. It therefore measures whether pairs that are more taxonomically distant also tend to be ordered as more distant in the learned representation.

B.3 DENDROGRAM–TAXONOMY AGREEMENT

The preceding metric operates directly on pairwise embedding distances. We additionally evaluate whether the hierarchical structure induced by those distances agrees with the known taxonomy.

For each evaluated representation, we construct an agglomerative hierarchical clustering using average linkage on the pairwise cosine distances. Let c_{ij} denote the cophenetic distance between samples i and j in the resulting dendrogram, i.e., the height at which the two samples are first merged into the same cluster. These distances are flattened into the condensed vector $\mathbf{c} = (c_{ij})_{i<j}$.

We define Dendrogram–Taxonomy Agreement as the Spearman correlation between the dendrogram cophenetic distances and the taxonomic distances:

$$\rho_{\mathrm{DT}};=;\rho_{\mathrm{S}}(\mathbf{c},\mathbf{v}). \tag{11}$$

This metric differs from the direct Spearman taxonomic distance correlation $\rho_{\mathrm{S}}(\mathbf{u},\mathbf{v})$. The latter measures agreement between the original embedding distances and taxonomic distances, whereas ρ_{DT} first converts the embedding geometry into a hierarchical clustering and then measures whether the relative separation represented by that dendrogram agrees with the taxonomic hierarchy. Higher values therefore indicate greater agreement between the hierarchy induced from the learned representation and the known taxonomic structure. We refer to this quantity as *Dendrogram–Taxonomy Agreement* rather than dendrogram purity, since it is a rank correlation between cophenetic and taxonomic distances rather than the conventional dendrogram-purity measure.

B.4 Protocol

For each evaluated representation, the two taxonomy-structure metrics are computed once over the corresponding test set and evaluated hierarchy. For each method, embeddings are extracted from the final trained encoder with its parameters fixed during evaluation. Pairwise embedding distances are cosine distances, and average linkage is used to construct the dendrogram for ρ_{DT}. All reported correlation coefficients are multiplied by 100 for readability.

C Datasets

C.1 iNaturalist 2021

We evaluate on the iNaturalist 2021 (iNat2021) dataset (Van Horn et al., 2021), a large-scale, fine-grained species classification dataset derived from user-submitted observations on the iNaturalist platform. iNat2021 spans the tree of life, covering 10,000 species across the kingdoms Animalia, Plantae, and Fungi, with each species annotated with its full Linnaean taxonomy (*kingdom, phylum, class, order, family, genus, species*). We use the *iNat2021 mini* train and validation sets. Because the official test split does not provide ground-truth labels, we use the validation set as the held-out test set for all reported evaluations.

Table 4 reports statistics for the nine iconic groups used in our experiments, which we refer to as evaluation groups. Plants and insects account for the majority of species and training images, while groups such as mollusks and amphibians are comparatively underrepresented. This imbalance motivates per-group evaluation, since group size and visual diversity may affect how well latent fine-grained taxonomic structure is preserved.

Table 4: Statistics for the nine iNaturalist 2021 iconic groups used in our experiments.

Group	Order	Family	Genus	Species	Train	Test
Plants	85	286	1,702	2,499	213,550	42,710
Insects	17	190	1,472	2,252	126,300	25,260
Birds	32	147	648	1,162	74,300	14,860
Fungi	36	102	210	323	17,050	3,410
Reptiles	4	40	158	296	15,650	3,130
Mammals	17	64	165	223	12,300	2,460
Fish	18	63	132	166	9,150	1,830
Amphibians	2	22	68	168	8,500	1,700
Mollusks	23	81	137	165	8,450	1,690

C.2 FGVC-Aircraft

We additionally evaluate on FGVC-Aircraft (Maji et al., 2013), a fine-grained aircraft recognition dataset with 10,000 images organized into a three-level hierarchy of *manufacturer*, *family*, and *variant*. The benchmark contains 30 manufacturers, 70 aircraft families, and 100 variants. Unlike

iNaturalist 2021, FGVC-Aircraft provides a compact non-biological hierarchy, allowing us to test whether TaxoCon generalizes beyond taxonomic structure.

We use *manufacturer* as the supervised coarse rank and withhold *family* and *variant* labels from representation fine-tuning and hyperparameter selection. The official training and validation partitions are combined for representation fine-tuning, while the official test partition is used for evaluation. Hyperparameters are transferred directly from iNaturalist 2021 without dataset-specific tuning. Table 5 summarizes the dataset hierarchy and split used in our experiments.

Table 5: FGVC-Aircraft statistics for the hierarchy and data splits used in our experiments.

Statistic	**Count**
Manufacturers	30
Families	70
Variants	100
Train + validation images	6,667
Test images	3,333

D TRAINING DETAILS

Backbones and initialization. All pretrained backbones operate on a fixed input resolution of 224×224 pixels unless otherwise specified. We consider three architectures with different inductive biases: *ResNet-50*, *ViT-S/16*, and *Swin-T*. *ResNet-50* and *Swin-T* are initialized from models pretrained on *ImageNet-1K* Deng et al. (2009), while *ViT-S/16* is initialized from a model pretrained on *ImageNet-21K* Ridnik et al. (2021) and subsequently fine-tuned on *ImageNet-1K*. We retain the pretrained weights as initialization and fine-tune the entire backbone on each target dataset rather than freezing any backbone layers.

Optimization and augmentation. For our primary baselines and TaxoCon, we use the same optimization protocol to ensure a controlled comparison across architectures. Models are trained using the *AdamW* optimizer with an initial learning rate of 1×10^{-4}. The learning rate follows a cosine decay schedule, preceded by a linear warm-up over the first *5%* of the total training steps. We train each model for a total of *50* epochs on each iNaturalist 2021 iconic group and on FGVC-Aircraft. For the single-view setting, each image contributes one view, and positive pairs are formed from distinct mini-batch samples sharing the same coarse training label; no additional augmented view of the same image is required. For TaxoCon(MV), two independently augmented views per image are encoded by the teacher and their pairwise similarities are averaged as described in Section 3. Training is performed using *bf16* mixed-precision arithmetic on a single 24 GB *NVIDIA A10G* GPU. Due to GPU memory limitations, we use a batch size of *300* for TaxoCon(MV) and MaskCon; a batch size of 800 for the other two TaxoCon variants—TaxoCon(SV) and TaxoCon w/o EMA—and for all baseline models.

MaskCon configuration. For MaskCon, we follow the original MaskCon implementation and augmentation protocol without modification. For both datasets, we use a batch size of *300*. All other MaskCon implementation and augmentation settings are kept identical to the original method, so that the comparison reflects the published MaskCon training procedure rather than adapting it to the TaxoCon protocol.

Representation extraction. The backbone produces a feature representation for each input image, which is subsequently ℓ_2-normalized before being used as the embedding for all downstream analyses. No additional projection head is applied when evaluating the learned representation; the normalized backbone feature itself is used for neighborhood-retrieval, linear-probing, and taxonomic-alignment analyses.

linear-probing. For linear-probing, we train a separate linear classifier for each taxonomic rank using only the training set and evaluate it on the held-out test set at the same rank. Labels from the

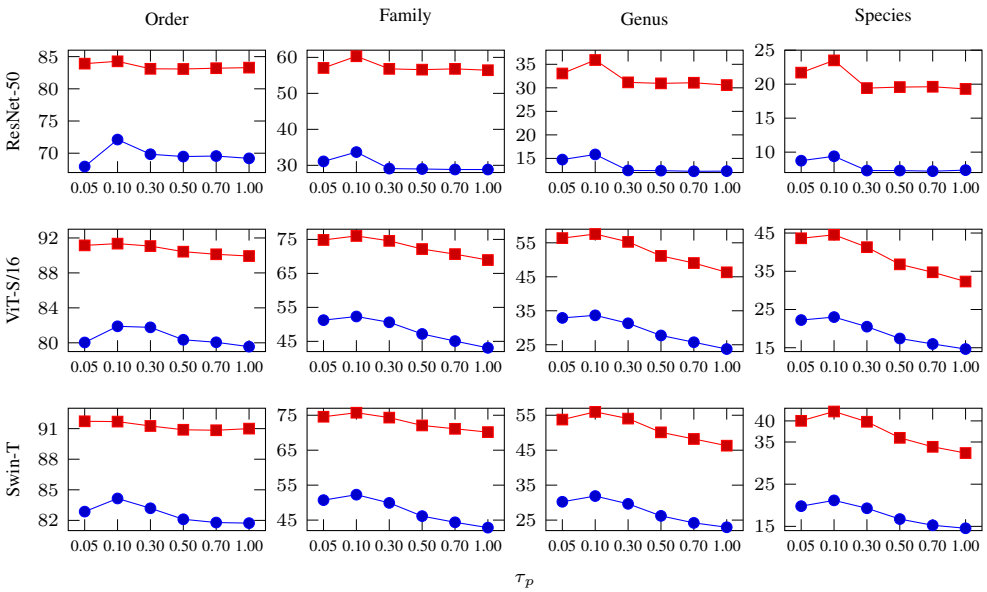

Figure 4: Recall@1 and Recall@5 macro-averaged across the nine iNaturalist groups across τ_p for three pretrained backbones and four taxonomic ranks. Blue curves denote Recall@1 and red curves denote Recall@5 for TaxoCon with EMA using $m = 0.95$.

held-out test set are never used for fitting the linear probe or selecting its parameters. For iNaturalist 2021, the evaluated ranks are *order*, *family*, *genus*, and *species*; for FGVC-Aircraft, they are *manufacturer*, *family*, and *variant*. This ensures that the reported linear-probing results measure the transferability of the learned representations to unseen test examples without test-set leakage.

E Ablation

E.1 Ablation results for τ_p

We examine the sensitivity of TaxoCon to the temperature parameter τ_p, which controls the sharpness of similarity-based positive weighting. We evaluate $\tau_p \in 0.05, 0.10, 0.30, 0.50, 0.70, 1.00$ while fixing the EMA momentum at $m = 0.95$. Figures 4–6 report results for Recall, linear-probing, and taxonomy-structure metrics.

Across backbones and evaluation criteria, $\tau_p = 0.10$ is generally optimal or close to optimal. The effect is most pronounced at the finer-grained ranks *family*, *genus*, and *species* levels. For ViT-S/16 and Swin-T, performance progressively declines as τ_p increases beyond 0.30, particularly at finer taxonomic ranks. This suggests that overly smooth positive weighting reduces TaxoCon's ability to emphasize the most informative same-label relationships.

The taxonomy-structure metrics show a similar pattern. Spearman and Dendrogram–Taxonomy Agreement typically improve from $\tau_p = 0.05$ to $\tau_p = 0.10$ and then remain stable or decline as τ_p increases. Thus, τ_p affects not only local retrieval and linear separability but also the global organization of the learned representation with respect to the underlying taxonomy.

Overall, TaxoCon exhibits a consistent operating point around $\tau_p = 0.10$ across backbones, taxonomic ranks, and evaluation protocols. These results motivate using $\tau_p = 0.10$ as the default configuration in the main experiments.

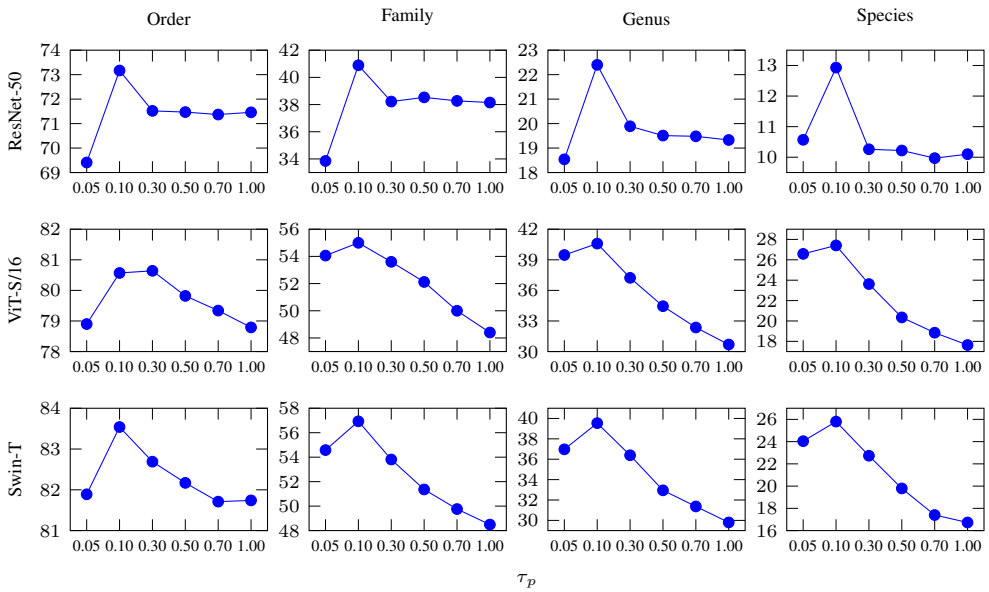

Figure 5: macro-averaged linear-probing top-1 accuracy across taxonomy levels for different τ_p values. Solid blue curves show TaxoCon with EMA using $m = 0.95$.

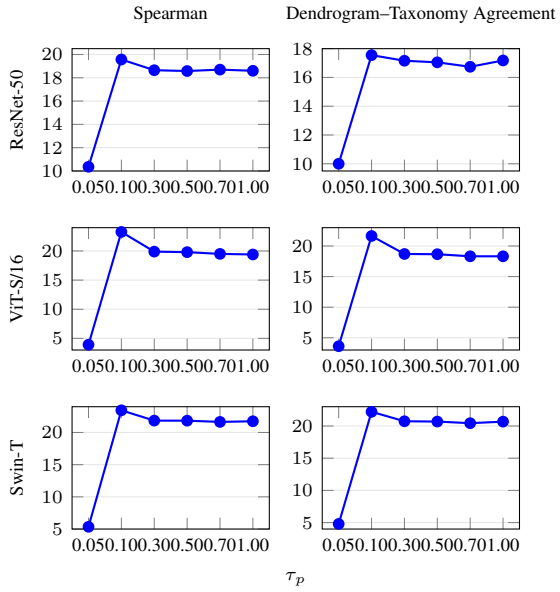

Figure 6: macro-averaged taxonomy-structure metrics across different τ_p values. Solid blue curves show TaxoCon with EMA using $m = 0.95$.

E.2 Ablation results for m

We investigate the sensitivity of TaxoCon to the EMA decay parameter m by fixing $\tau_p = 0.10$ and evaluating $m \in 0.900, 0.95, 0.999$. Figures 7–9 report neighborhood-retrieval, linear-probing, and taxonomy-structure performance across the three backbone architectures.

Overall, TaxoCon is relatively robust to m within the tested range, although the preferred value varies slightly across backbones and metrics. For ResNet-50, $m = 0.90$ often provides the strongest Recall and linear-probing performance, with $m = 0.95$ remaining close. ViT-S/16 generally favors $m = 0.95$ at the finer-grained ranks, while Swin-T shows the clearest preference for $m = 0.95$

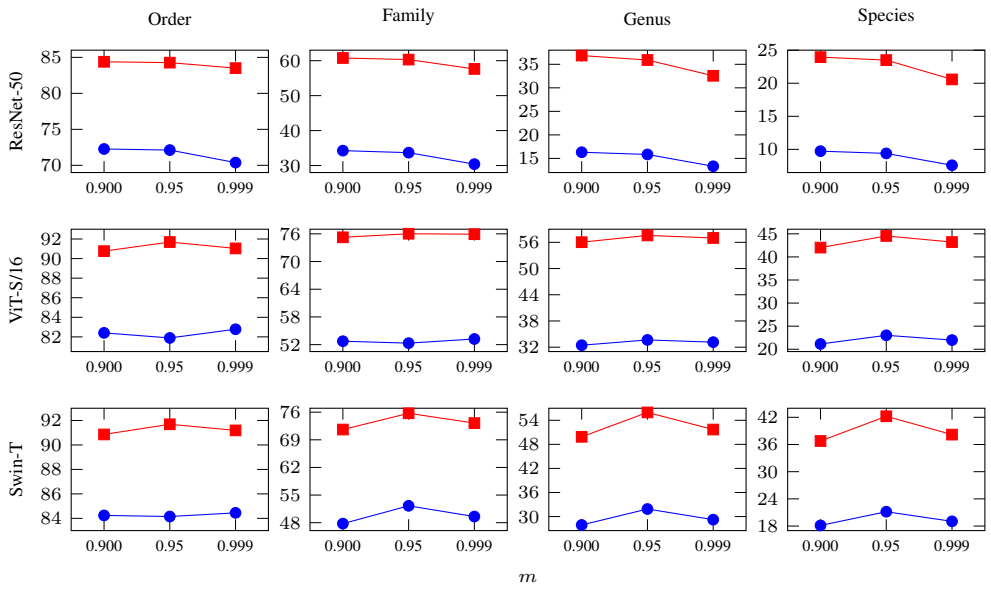

Figure 7: macro-averaged Recall@1 and Recall@5 performance across EMA decay m for three pretrained backbones and four taxonomic ranks. Blue curves denote Recall@1 and red curves denote Recall@5 for TaxoCon with $\tau_p = 0.10$ fixed.

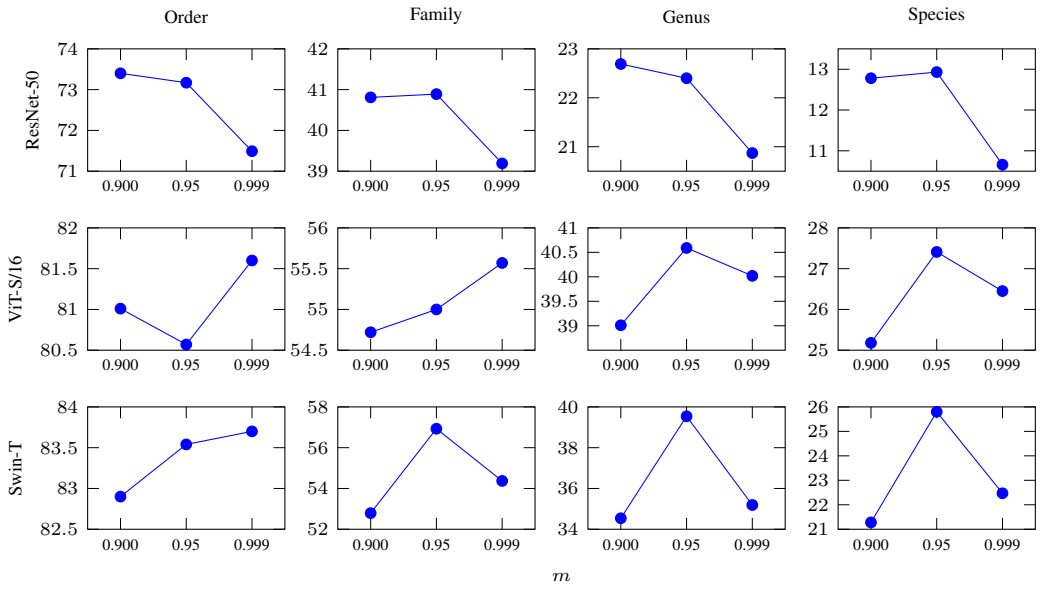

Figure 8: macro-averaged linear-probing accuracy across taxonomy levels for different EMA decay m values, with $\tau_p = 0.10$ fixed for TaxoCon with EMA. Solid blue curves with circle markers show TaxoCon.

across Recall and linear-probing evaluations. The taxonomy-structure metrics similarly show strong or near-best performance at $m = 0.95$, whereas increasing the decay to $m = 0.999$ provides no consistent improvement.

These results indicate that TaxoCon is not highly sensitive to the precise EMA decay within the evaluated range. Overall, $m = 0.95$ provides the most consistent performance across backbones, taxonomic ranks, and evaluation protocols, motivating its use as the default EMA decay.

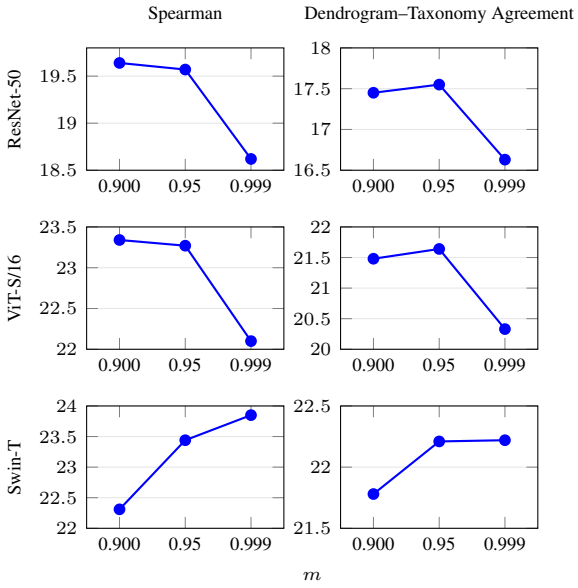

Figure 9: macro-averaged taxonomy-structure metrics across EMA decay m values, with $\tau_p = 0.10$ fixed for TaxoCon. Blue solid curves with circle markers show TaxoCon with EMA.

F PER-GROUP RESULTS

F.1 RECALL RESULTS

Metric	Method	Mollusks	Mammals	Fishes	Insects	Birds	Reptiles	Amphibians	Fungi	Plants
Order	Frozen Backbone	59.76 / 82.25	70.85 / 90.85	71.04 / 89.62	79.83 / 93.09	69.30 / 87.31	95.72 / **99.01**	91.24 / **98.88**	62.76 / 87.07	43.51 / 69.90
	SupCon	69.29 / 85.09	72.64 / 88.05	77.21 / 87.70	88.69 / 94.98	77.86 / 88.38	96.07 / 98.15	94.00 / 97.88	71.85 / 88.04	60.09 / 76.14
	MS-loss	67.34 / 85.38	76.54 / 86.02	77.70 / 87.60	88.38 / 93.70	79.43 / 87.56	97.25 / 98.66	94.94 / 97.53	75.60 / 88.30	64.78 / 76.57
	Cross Entropy	55.03 / 74.56	50.81 / 78.94	55.30 / 78.58	86.87 / 93.44	71.35 / 85.16	90.93 / 97.03	93.29 / 96.76	51.79 / 77.74	46.90 / 65.56
	TaxoCon w/o EMA	69.76 / 84.08	74.39 / 86.50	80.60 / 89.67	88.19 / 94.11	76.88 / 87.79	96.33 / 98.31	94.24 / 97.88	74.96 / 87.62	60.58 / 75.88
	TaxoCon(SV)	75.03 / **87.99**	78.90 / 91.22	82.02 / 91.69	88.73 / 94.85	78.16 / 89.64	97.22 / 98.88	94.35 / 98.24	78.59 / 91.14	64.02 / 78.50
	TaxoCon(MV)	**80.53** / 87.75	**81.30 / 91.42**	**82.84 / 91.91**	**92.31 / 95.67**	**82.24 / 90.90**	**97.76 / 99.01**	**95.94** / 98.12	**86.57 / 92.82**	**65.42 / 80.12**
Family	Frozen Backbone	37.10 / 64.08	56.99 / 79.76	44.92 / 72.57	48.61 / 73.51	43.21 / 65.79	48.75 / 66.81	44.76 / 72.18	43.23 / 68.79	38.33 / 62.41
	SupCon	37.75 / 64.38	44.72 / 67.68	33.99 / 60.33	49.16 / 75.05	40.71 / 60.87	29.23 / 62.04	38.00 / 74.00	39.79 / 66.89	46.66 / 65.11
	MS-loss	42.78 / 70.30	55.04 / 72.48	43.39 / 67.54	50.25 / 75.67	43.76 / 62.11	44.63 / 72.59	46.65 / 75.88	49.91 / 73.99	54.11 / 69.34
	Cross Entropy	19.88 / 46.63	24.67 / 51.30	15.03 / 40.05	26.57 / 58.05	27.37 / 46.40	20.29 / 52.62	30.71 / 66.59	19.65 / 44.46	32.40 / 51.69
	TaxoCon w/o EMA	37.51 / 64.73	46.95 / 68.70	41.97 / 67.87	41.79 / 70.06	38.31 / 58.08	30.32 / 61.85	41.24 / 75.12	42.64 / 66.92	46.93 / 65.20
	TaxoCon(SV)	47.57 / 71.95	62.44 / 81.99	53.17 / 76.28	52.72 / 77.21	50.26 / 70.22	49.27 / **77.22**	45.06 / **78.53**	55.43 / 78.97	54.98 / 71.82
	TaxoCon(MV)	**55.09 / 77.04**	**65.57 / 83.33**	**55.68 / 77.98**	**57.38 / 80.09**	**55.42 / 73.16**	**50.58** / 77.13	**48.00** / 77.59	**64.78 / 83.55**	**58.94 / 75.10**
Genus	Frozen Backbone	31.54 / 54.56	27.07 / 53.09	37.21 / 61.37	22.51 / 43.25	22.12 / 41.51	20.50 / 42.73	21.94 / 49.82	35.31 / 62.35	24.73 / 43.16
	SupCon	27.93 / 51.30	20.53 / 41.71	21.37 / 45.46	21.55 / 41.75	19.49 / 38.78	10.54 / 29.17	17.12 / 44.12	28.42 / 53.40	23.49 / 40.12
	MS-loss	35.62 / 60.00	30.37 / 50.89	32.51 / 55.90	21.21 / 41.39	23.06 / 42.40	21.85 / 45.59	23.00 / 49.53	39.62 / 62.93	32.56 / 49.48
	Cross Entropy	13.14 / 32.96	8.21 / 23.74	7.60 / 22.24	6.50 / 16.39	8.65 / 22.30	5.69 / 17.25	11.29 / 34.76	11.67 / 29.94	12.09 / 25.03
	TaxoCon w/o EMA	27.75 / 51.66	22.64 / 43.29	27.54 / 52.57	15.35 / 32.29	16.86 / 35.79	11.18 / 29.01	19.18 / 46.00	30.26 / 53.43	23.52 / 40.23
	TaxoCon(SV)	37.46 / 63.02	40.73 / 65.61	42.68 / 66.50	23.40 / 44.79	30.32 / 51.82	25.65 / 51.57	22.12 / 50.53	45.66 / 70.56	34.88 / 53.84
	TaxoCon(MV)	**44.26 / 68.46**	**42.76 / 68.50**	**44.97 / 68.74**	**27.21 / 49.22**	**34.47 / 56.45**	**27.70 / 53.10**	**24.53 / 52.24**	**55.19 / 76.66**	**36.50 / 56.13**
Species	Frozen Backbone	28.22 / 50.83	22.13 / 46.46	29.62 / 53.01	15.05 / 31.47	10.80 / 24.67	9.43 / 24.05	10.35 / 27.35	26.39 / 52.29	10.83 / 24.56
	SupCon	23.49 / 45.62	13.33 / 31.26	14.92 / 37.21	13.84 / 29.51	9.01 / 21.39	4.95 / 14.06	8.18 / 19.71	19.74 / 42.05	9.35 / 22.37
	MS-loss	31.72 / 56.21	21.18 / 40.28	24.92 / 47.27	13.44 / 28.96	11.31 / 25.03	10.26 / 26.81	10.35 / 26.00	28.71 / 52.46	16.53 / 33.07
	Cross Entropy	10.12 / 27.57	4.43 / 14.07	4.70 / 15.14	3.04 / 8.58	3.16 / 9.89	2.20 / 6.96	4.06 / 13.59	7.39 / 19.38	4.18 / 11.61
	TaxoCon w/o EMA	23.43 / 47.40	14.11 / 31.95	19.51 / 43.06	9.18 / 20.75	7.10 / 19.09	5.18 / 13.55	8.00 / 22.24	21.32 / 41.67	9.31 / 22.38
	TaxoCon(SV)	32.60 / 59.11	30.77 / **55.65**	34.10 / 59.13	15.38 / 32.30	16.83 / 35.24	12.27 / 31.63	10.53 / 28.00	35.89 / 61.58	18.80 / 38.19
	TaxoCon(MV)	**38.64 / 64.85**	**32.28** / 52.66	**36.83 / 62.62**	**17.97 / 36.62**	**19.83 / 39.99**	**13.96 / 33.45**	**13.24 / 29.06**	**43.90 / 69.50**	**20.12 / 41.32**

Table 6: Per-group Recall@1 / Recall@5 for ViT-S/16.

Metric	Method	Mollusks	Mammals	Fishes	Insects	Birds	Reptiles	Amphibians	Fungi	Plants
Order	Frozen Backbone	62.49 / 82.37	71.54 / 90.00	66.78 / 87.98	80.99 / 93.73	68.51 / 87.06	96.93 / 99.04	92.82 / 95.18	61.32 / 85.34	26.82 / 55.83
	SupCon	75.27 / 87.04	76.26 / 89.39	79.89 / 90.27	90.13 / 96.20	78.50 / 89.86	96.84 / 99.14	95.06 / 97.82	71.85 / 88.04	63.09 / 78.03
	MS-loss	76.98 / 88.22	78.86 / 87.07	81.26 / 88.47	91.75 / 95.49	80.98 / 89.03	97.16 / 98.79	96.47 / 98.41	75.60 / 88.30	62.96 / 76.24
	Cross Entropy	69.88 / 83.61	68.46 / 85.37	71.80 / 85.57	91.08 / 96.29	78.53 / 88.88	96.42 / 98.37	95.18 / 97.18	51.79 / 77.74	56.50 / 74.33
	TaxoCon w/o EMA	77.69 / 87.28	77.89 / 88.21	81.97 / 89.62	90.97 / 95.66	80.26 / 88.99	97.00 / 98.79	94.12 / 97.41	74.96 / 87.62	65.51 / 78.63
	TaxoCon(SV)	78.82 / **88.99**	82.20 / **91.99**	83.01 / **91.97**	92.87 / **95.87**	83.10 / 89.97	**97.89 / 99.20**	95.82 / **98.29**	78.59 / 91.14	65.06 / 77.75
	TaxoCon(MV)	**81.30** / 88.58	**83.37** / 91.14	**85.96** / 91.20	**92.93 / 95.87**	**85.42 / 90.49**	96.87 / 99.04	**96.06** / 97.76	**87.18 / 92.55**	**76.63 / 84.89**
Family	Frozen Backbone	35.62 / 64.08	57.07 / 73.76	32.08 / 61.91	46.47 / 70.62	36.97 / 61.39	43.99 / 72.04	41.88 / 73.65	39.97 / 69.09	21.51 / 45.25
	SupCon	30.59 / 58.46	44.76 / 72.24	43.06 / 69.13	35.60 / 68.93	38.20 / 63.01	37.73 / 69.07	31.82 / 71.00	39.79 / 66.89	50.47 / 68.64
	MS-loss	47.22 / 72.19	56.34 / 75.73	41.04 / 65.14	53.35 / 78.29	43.22 / 62.72	44.22 / 74.54	45.47 / 79.00	49.91 / 73.99	50.73 / 67.88
	Cross Entropy	30.47 / 58.64	39.80 / 64.84	23.99 / 52.19	44.71 / 72.34	36.40 / 56.57	31.41 / 63.07	41.35 / 73.29	19.65 / 44.46	39.94 / 60.88
	TaxoCon w/o EMA	45.56 / 71.01	52.24 / 74.72	44.86 / 67.60	41.62 / 69.70	40.39 / 60.05	33.77 / 66.04	41.00 / 75.59	42.64 / 66.92	49.22 / 67.27
	TaxoCon(SV)	49.17 / 74.56	64.55 / **82.93**	44.75 / 71.69	53.17 / 78.15	47.20 / 67.40	**54.73 / 80.80**	**48.94** / 77.24	55.43 / 78.97	52.52 / 69.71
	TaxoCon(MV)	**51.78 / 75.03**	**66.02** / 82.15	**50.05 / 73.11**	**55.23 / 78.84**	**51.09 / 68.79**	51.76 / 80.19	48.06 / **78.47**	**63.17 / 82.61**	**65.42 / 78.84**
Genus	Frozen Backbone	28.88 / 54.38	37.68 / 62.20	24.54 / 49.51	14.29 / 30.11	21.33 / 41.57	25.72 / 54.31	15.71 / 44.53	32.67 / 58.68	12.41 / 27.01
	SupCon	20.00 / 45.03	18.82 / 43.50	32.62 / 56.99	10.04 / 25.52	15.14 / 36.90	15.85 / 38.43	10.41 / 34.53	28.42 / 53.40	27.89 / 46.26
	MS-loss	37.22 / 62.78	31.99 / 54.96	30.05 / 51.69	21.74 / 42.59	22.15 / 42.36	20.73 / 45.88	21.53 / 50.59	39.62 / 62.93	28.41 / 46.50
	Cross Entropy	20.47 / 45.80	16.26 / 36.67	14.64 / 35.25	15.66 / 33.42	14.41 / 32.31	10.54 / 29.17	18.18 / 44.06	11.67 / 29.94	16.00 / 31.38
	TaxoCon w/o EMA	34.73 / 60.36	26.99 / 51.10	33.83 / 55.79	14.01 / 29.76	17.88 / 37.49	11.79 / 31.18	18.12 / 46.41	30.26 / 53.43	23.85 / 40.70
	TaxoCon(SV)	39.23 / 64.97	42.72 / **67.32**	34.37 / 58.91	19.67 / 40.94	24.44 / 47.10	**28.75 / 56.65**	**24.59** / 50.41	45.66 / 70.56	27.48 / 46.79
	TaxoCon(MV)	**40.65 / 66.57**	**42.85** / 67.28	**38.09 / 61.26**	**22.02 / 43.32**	**27.95 / 49.85**	25.78 / 55.05	22.88 / **51.47**	**52.11 / 74.46**	**42.41 / 61.19**
Species	Frozen Backbone	24.85 / 49.59	13.90 / 33.28	18.52 / 40.00	12.02 / 20.82	11.28 / 26.14	8.52 / 24.54	5.06 / 14.29	23.96 / 48.18	7.42 / 16.89
	SupCon	16.57 / 40.41	12.85 / 31.59	25.14 / 50.22	5.95 / 15.27	7.03 / 19.14	7.64 / 20.93	3.71 / 13.24	19.74 / 42.05	13.75 / 29.89
	MS-loss	32.25 / 57.51	22.76 / 44.63	22.24 / 44.04	13.41 / 28.78	10.81 / 24.91	10.61 / 26.49	9.06 / 26.00	28.71 / 52.46	14.81 / 30.01
	Cross Entropy	15.56 / 39.17	9.72 / 26.34	8.80 / 25.68	9.05 / 21.24	5.65 / 15.46	4.47 / 14.06	6.94 / 20.29	7.39 / 19.38	6.01 / 15.46
	TaxoCon w/o EMA	28.88 / 55.50	18.58 / 40.20	26.01 / 47.98	7.59 / 18.11	8.07 / 20.50	5.05 / 15.43	7.06 / 20.82	21.32 / 41.67	10.50 / 23.68
	TaxoCon(SV)	33.37 / 60.18	31.91 / **57.44**	26.78 / 50.87	11.57 / 27.29	12.15 / 28.94	**14.22 / 36.39**	**11.65 / 28.18**	35.89 / 61.58	12.93 / 29.19
	TaxoCon(MV)	**34.97 / 61.83**	**32.40** / 57.07	**29.78 / 53.06**	**13.58 / 29.77**	**14.70 / 32.02**	12.52 / 34.70	11.06 / 27.47	**39.56 / 65.63**	**24.53 / 44.89**

Table 7: per-group Recall@1 / Recall@5 for Swin-T.

Metric	Method	Mollusks	Mammals	Fishes	Insects	Birds	Reptiles	Amphibians	Fungi	Plants
Order	Frozen Backbone	54.44 / 79.05	67.64 / **89.19**	63.28 / 80.03	76.96 / 91.88	64.15 / 81.13	92.75 / 94.79	90.71 / **99.24**	52.38 / 80.79	22.50 / 51.12
	SupCon	61.30 / 80.41	65.98 / 84.15	68.63 / 84.21	85.05 / 92.82	69.16 / 83.56	95.34 / 97.99	93.65 / 97.76	59.68 / 79.82	25.64 / 47.69
	MS-loss	56.98 / 75.98	66.10 / 81.59	65.85 / 78.20	85.31 / 91.70	69.11 / 81.65	95.56 / 97.16	93.00 / 96.41	58.15 / 76.16	23.31 / 45.30
	MaskCon	67.81 / 84.02	71.38 / 87.52	70.87 / 87.27	88.22 / 94.73	76.91 / **88.53**	94.23 / 98.40	93.88 / 97.94	73.58 / **88.27**	47.48 / 69.55
	TaxoCon w/o EMA	63.73 / 80.06	68.54 / 83.58	69.45 / 84.26	85.44 / 91.97	69.92 / 82.94	95.30 / 97.76	94.41 / 96.41	62.99 / 81.06	29.53 / 49.74
	TaxoCon(SV)	64.73 / 82.19	70.98 / 84.80	70.49 / 84.81	85.96 / 92.93	71.40 / 84.22	94.86 / 97.76	94.00 / 97.06	64.46 / 81.94	32.79 / 53.15
	TaxoCon(MV)	**70.83 / 85.50**	**76.06** / 88.82	**74.15 / 87.81**	**90.59 / 94.97**	**78.41 / 88.53**	**96.10 / 98.50**	**95.65** / 97.59	**73.75** / 87.95	**54.16 / 70.59**
Family	Frozen Backbone	26.92 / 56.57	42.67 / 64.23	25.57 / 56.72	33.01 / 62.73	27.16 / 46.33	31.26 / 65.50	32.92 / 65.10	27.78 / 54.00	17.46 / 32.45
	SupCon	27.40 / 54.26	41.75 / 65.61	26.50 / 53.06	32.93 / 63.00	26.85 / 47.40	29.65 / 64.12	31.76 / 70.06	26.77 / 53.72	16.22 / 33.26
	MS-loss	23.20 / 50.35	43.37 / 64.39	20.82 / 45.63	36.53 / 65.56	27.97 / 48.73	31.41 / 63.67	33.47 / 71.88	26.98 / 52.79	15.95 / 33.17
	Cross Entropy	17.28 / 42.25	33.46 / 55.69	13.83 / 36.61	21.31 / 50.13	21.46 / 40.80	18.91 / 49.17	26.71 / 64.12	15.31 / 38.15	11.58 / 27.00
	MaskCon	**39.47** / 66.15	48.70 / 73.86	33.01 / 61.42	43.52 / 71.53	37.49 / 58.58	32.49 / 65.65	25.59 / 64.47	**45.66** / 71.96	38.19 / 60.53
	TaxoCon w/o EMA	29.94 / 56.98	42.11 / 67.20	26.99 / 54.97	29.94 / 58.88	26.94 / 46.92	29.68 / 63.87	33.82 / 70.00	29.09 / 56.30	18.88 / 35.79
	TaxoCon(SV)	33.02 / 62.78	49.47 / 71.14	29.34 / 56.28	37.65 / 65.55	30.70 / 50.75	35.02 / 68.05	34.71 / 70.65	31.94 / 58.56	21.44 / 39.65
	TaxoCon(MV)	37.63 / **66.57**	**54.88 / 76.67**	**35.85 / 64.21**	**48.11 / 73.63**	**40.64 / 61.08**	**36.33 / 68.53**	**35.29 / 72.53**	44.05 / **72.20**	**42.04 / 61.33**
Genus	Frozen Backbone	20.06 / 45.62	21.34 / 42.90	17.76 / 35.06	7.69 / 22.40	10.02 / 21.35	10.80 / 31.31	**15.82 / 42.76**	17.22 / 38.30	4.04 / 11.45
	SupCon	18.99 / 42.43	18.09 / 40.53	16.50 / 36.94	8.71 / 21.99	8.46 / 22.53	9.27 / 27.76	10.53 / 33.94	16.63 / 38.42	3.46 / 10.30
	MS-loss	17.10 / 38.58	18.25 / 39.84	13.99 / 30.16	8.95 / 22.91	10.24 / 24.91	10.38 / 28.63	12.82 / 35.41	17.60 / 38.86	4.66 / 12.77
	Cross Entropy	10.47 / 28.46	10.81 / 26.95	7.43 / 21.97	3.31 / 9.89	5.05 / 15.92	3.83 / 14.06	7.71 / 26.47	8.65 / 23.05	1.86 / 6.35
	MaskCon	**31.18 / 56.39**	24.43 / 48.33	23.33 / 46.56	14.66 / 32.56	16.49 / 36.14	13.16 / 32.27	8.47 / 30.47	**36.16** / 60.76	19.73 / **39.17**
	TaxoCon w/o EMA	20.00 / 44.20	17.97 / 41.18	17.16 / 40.77	5.96 / 17.03	8.26 / 22.36	10.42 / 28.91	11.41 / 35.29	18.12 / 41.35	4.38 / 11.76
	TaxoCon(SV)	23.79 / 49.59	24.39 / 47.40	19.84 / 41.26	9.70 / 24.68	10.69 / 26.73	14.41 / 36.04	12.88 / 38.76	21.11 / 44.49	5.61 / 14.64
	TaxoCon(MV)	27.57 / 54.67	**29.59 / 55.77**	**24.86 / 49.45**	**17.96 / 37.07**	**17.52 / 38.20**	**15.43 / 38.02**	13.94 / 40.71	34.22 / **61.14**	**20.66** / 38.68
Species	Frozen Backbone	16.80 / 40.41	13.39 / 30.93	13.83 / 31.10	4.56 / 12.50	4.87 / 12.32	3.99 / 13.77	3.46 / 10.74	10.04 / 30.22	1.15 / 3.24
	SupCon	14.91 / 35.92	11.63 / 29.47	12.02 / 29.78	4.54 / 12.10	2.84 / 8.91	3.61 / 12.56	3.65 / 12.18	11.00 / 28.01	0.85 / 3.13
	MS-loss	14.08 / 33.61	10.85 / 28.90	10.33 / 30.16	4.39 / 12.56	3.90 / 11.00	4.06 / 12.59	3.82 / 13.47	11.17 / 27.86	1.94 / 5.85
	Cross Entropy	7.46 / 22.72	5.93 / 17.03	4.75 / 14.86	1.16 / 4.16	1.44 / 5.55	1.31 / 4.73	1.53 / 7.76	4.37 / 13.93	0.54 / 2.03
	MaskCon	**26.98 / 52.01**	16.59 / 37.76	15.47 / 38.42	8.25 / 20.36	7.44 / 19.14	5.78 / 15.72	2.35 / 10.06	**27.16 / 50.50**	8.88 / **22.83**
	TaxoCon w/o EMA	16.57 / 38.76	11.14 / 30.24	12.62 / 31.48	2.87 / 8.32	2.65 / 8.76	4.12 / 12.52	4.06 / 12.94	12.26 / 30.06	1.07 / 3.77
	TaxoCon(SV)	19.70 / 43.96	16.46 / 37.11	14.54 / 34.43	4.98 / 13.75	3.42 / 11.45	5.62 / 17.67	4.12 / 15.06	14.08 / 33.02	1.40 / 5.06
	TaxoCon(MV)	23.49 / 49.17	**20.28 / 43.25**	**17.16 / 40.71**	**10.76 / 23.96**	**7.54 / 20.29**	**7.28 / 19.71**	**4.76 / 15.53**	26.67 / **50.50**	**9.67** / 22.75

Table 8: per-group Recall@1 / Recall@5 for ResNet-50.

F.2 LINEAR-PROBING RESULTS

Metric	Method	Mollusks	Mammals	Fishes	Insects	Birds	Reptiles	Amphibians	Fungi	Plants
Order	Frozen Backbone	66.86	68.90	71.86	82.72	65.01	96.81	90.59	60.56	45.54
	SupCon	67.16	67.68	77.05	89.31	71.06	96.65	95.29	71.41	64.65
	Cross Entropy	52.07	53.25	59.29	87.71	72.14	92.17	90.59	53.08	54.11
	MS-loss	68.93	71.34	74.59	87.19	70.05	97.44	95.29	71.99	63.86
	TaxoCon w/o EMA	69.53	70.53	79.51	89.81	69.31	95.69	94.12	73.75	64.13
	TaxoCon(SV)	73.96	**79.27**	**81.97**	89.09	71.97	**98.24**	**95.59**	79.47	64.82
	TaxoCon(MV)	**78.99**	78.05	81.69	**91.55**	**78.97**	97.28	95.29	**87.24**	**65.49**
Family	Frozen Backbone	47.93	62.20	52.73	48.81	43.10	53.35	**56.76**	50.59	38.39
	SupCon	37.87	49.19	42.35	55.68	43.10	36.42	45.59	48.68	50.78
	Cross Entropy	27.51	28.46	22.13	41.61	34.05	25.56	35.29	25.81	39.79
	MS-loss	48.82	55.89	44.26	52.99	41.89	48.40	49.12	50.00	50.74
	TaxoCon w/o EMA	43.49	48.58	45.36	55.56	41.49	35.14	49.12	46.92	51.60
	TaxoCon(SV)	53.55	61.18	57.92	57.22	47.98	55.43	54.41	60.70	51.70
	TaxoCon(MV)	**59.76**	**65.45**	**59.84**	**59.07**	**52.99**	**55.91**	52.06	**64.52**	**52.20**
Genus	Frozen Backbone	39.64	44.31	45.36	32.11	33.24	33.87	**30.88**	47.95	33.15
	SupCon	33.43	27.24	27.60	29.30	26.45	18.05	25.29	36.80	30.31
	Cross Entropy	18.05	11.79	13.11	12.29	14.30	10.22	16.18	17.30	17.96
	MS-loss	43.49	32.93	35.52	28.84	27.52	30.19	26.18	40.32	35.40
	TaxoCon w/o EMA	33.14	27.44	33.33	29.59	25.40	15.50	25.59	34.16	38.02
	TaxoCon(SV)	48.22	44.92	**47.54**	31.59	33.38	**37.06**	**30.88**	52.93	33.63
	TaxoCon(MV)	**50.30**	**45.73**	**47.54**	**35.21**	**39.54**	34.67	29.12	**57.63**	**40.93**
Species	Frozen Backbone	35.80	**35.16**	33.88	0.06	20.19	**18.85**	13.82	37.68	0.07
	SupCon	28.40	18.50	18.03	16.84	9.99	5.59	6.76	24.78	10.35
	Cross Entropy	14.20	4.27	7.92	4.77	4.10	3.35	5.00	10.26	4.88
	MS-loss	39.94	23.78	25.14	15.60	11.57	13.26	13.24	32.70	13.46
	TaxoCon w/o EMA	29.29	16.87	20.77	16.86	10.20	4.47	7.65	26.25	15.91
	TaxoCon(SV)	43.49	31.30	**38.52**	19.89	14.10	17.25	**16.18**	43.84	13.43
	TaxoCon(MV)	**44.97**	32.32	38.25	**21.52**	**20.52**	16.13	15.88	**49.27**	**17.22**

Table 9: per-group linear-probing accuracy for ViT-S/16

Metric	Method	Mollusks	Mammals	Fishes	Insects	Birds	Reptiles	Amphibians	Fungi	Plants
Order	Frozen Backbone	65.38	73.37	69.95	80.72	66.66	98.24	93.82	65.10	32.69
	SupCon	78.11	75.81	80.87	88.86	71.06	97.92	94.71	79.91	62.71
	Cross Entropy	71.01	69.51	70.49	89.31	72.14	96.17	95.29	75.81	59.63
	MS-loss	78.11	76.02	80.05	88.64	78.20	98.08	96.76	71.99	60.08
	TaxoCon w/o EMA	75.15	72.56	80.97	88.54	69.31	98.40	95.59	82.40	64.47
	TaxoCon(SV)	79.88	80.49	84.97	89.39	71.97	**99.04**	95.88	82.40	65.75
	TaxoCon(MV)	**81.36**	**82.72**	**85.25**	**90.93**	**83.34**	97.28	**96.76**	**88.27**	**70.92**
Family	Frozen Backbone	46.75	61.23	45.55	51.56	41.99	57.99	53.53	51.47	24.40
	SupCon	39.64	50.20	43.44	48.83	43.10	45.53	47.06	46.33	48.08
	Cross Entropy	38.76	46.54	34.97	59.34	34.05	47.12	49.12	47.36	47.08
	MS-loss	55.03	57.32	49.18	58.33	45.22	52.40	51.18	50.00	43.23
	TaxoCon w/o EMA	44.97	50.20	46.45	48.71	41.49	45.53	44.12	53.52	49.18
	TaxoCon(SV)	52.96	63.01	**55.74**	52.24	47.98	**55.43**	52.35	60.41	52.03
	TaxoCon(MV)	**58.28**	**68.09**	55.46	**58.63**	**52.99**	54.79	**53.53**	**67.60**	**63.36**
Genus	Frozen Backbone	37.87	40.02	35.97	30.64	28.30	30.40	35.29	47.95	24.26
	SupCon	28.11	23.78	33.61	19.52	26.45	21.57	21.18	34.90	34.77
	Cross Entropy	29.88	24.19	23.77	27.83	14.30	21.73	0.88	35.63	28.69
	MS-loss	41.85	36.66	33.06	30.92	27.25	27.16	28.82	40.32	32.19
	TaxoCon w/o EMA	35.80	27.03	33.61	17.89	25.40	20.29	22.06	41.50	34.93
	TaxoCon(SV)	42.01	41.87	**44.26**	17.54	**33.38**	31.47	25.29	51.17	38.11
	TaxoCon(MV)	**46.75**	**46.14**	42.35	**32.01**	33.34	**33.71**	**30.29**	**56.60**	**47.59**
Species	Frozen Backbone	34.32	34.55	23.50	6.60	16.25	12.19	10.51	37.39	4.24
	SupCon	21.89	16.46	23.22	8.27	9.99	8.63	5.00	22.43	14.58
	Cross Entropy	23.37	11.79	15.03	14.92	4.10	6.23	9.41	22.58	11.40
	MS-loss	34.45	23.58	25.96	16.47	11.34	11.34	11.76	32.70	15.94
	TaxoCon w/o EMA	25.44	17.68	25.14	6.69	10.20	7.19	4.41	27.71	15.78
	TaxoCon(SV)	35.80	29.47	33.61	7.03	14.10	13.90	11.76	39.59	17.89
	TaxoCon(MV)	**42.60**	**31.91**	**35.79**	**18.37**	**14.54**	**17.09**	**14.41**	**44.87**	**24.91**

Table 10: per-group linear-probing accuracy for Swin-T.

Metric	Method	Mollusks	Mammals	Fishes	Insects	Birds	Reptiles	Amphibians	Fungi	Plants
Order	Frozen Backbone	62.33	71.15	68.19	78.89	67.27	97.21	93.43	60.61	24.80
	SupCon	65.68	68.29	69.40	82.70	68.44	97.12	95.88	61.88	27.10
	Cross Entropy	61.24	61.79	62.30	83.45	66.25	94.09	96.18	54.40	22.61
	MS-loss	61.24	68.50	67.21	81.93	66.79	96.65	96.47	56.30	20.28
	MaskCon	69.53	73.98	72.95	85.99	76.58	96.81	95.59	72.87	41.59
	TaxoCon w/o EMA	**70.41**	69.51	72.13	83.08	68.41	96.96	95.59	63.05	28.51
	TaxoCon(SV)	69.82	70.93	74.86	83.06	69.01	97.44	96.47	65.10	31.78
	TaxoCon(MV)	70.12	**77.64**	**76.78**	**88.32**	**77.46**	**98.08**	**97.06**	**74.49**	**46.27**
Family	Frozen Backbone	42.49	58.53	36.20	48.22	38.92	**55.66**	**51.66**	42.27	19.71
	SupCon	40.83	49.59	34.97	41.88	33.92	45.21	42.94	36.36	19.14
	Cross Entropy	27.81	39.23	22.40	36.10	25.71	26.84	34.41	23.17	13.85
	MS-loss	35.80	50.61	31.15	45.19	33.78	47.76	45.29	31.38	16.21
	MaskCon	**46.75**	56.50	38.80	53.42	42.29	46.01	39.12	47.36	35.52
	TaxoCon w/o EMA	39.05	51.63	36.89	41.67	32.07	47.12	40.88	36.22	20.21
	TaxoCon(SV)	44.38	53.05	35.52	45.29	33.92	49.68	45.00	38.71	22.44
	TaxoCon(MV)	45.56	**59.15**	**45.08**	**53.86**	**45.73**	47.12	46.76	**48.97**	**40.58**
Genus	Frozen Backbone	29.55	28.54	23.05	13.80	15.01	**26.36**	19.43	31.09	8.84
	SupCon	28.70	25.41	22.40	14.19	14.20	21.88	15.29	26.10	8.59
	Cross Entropy	12.13	15.45	12.84	6.81	9.19	6.07	12.65	14.22	4.61
	MS-loss	27.63	28.25	23.36	15.09	15.02	24.12	17.65	26.98	7.81
	MaskCon	**38.46**	31.50	**28.42**	25.36	23.62	21.73	14.12	39.44	24.88
	TaxoCon w/o EMA	28.11	27.64	23.22	14.51	14.37	20.77	15.59	26.83	8.15
	TaxoCon(SV)	31.07	29.88	27.05	15.66	15.68	23.64	20.29	27.71	10.65
	TaxoCon(MV)	35.80	**34.55**	27.60	**25.77**	**23.72**	23.16	**22.35**	**40.18**	**27.04**
Species	Frozen Backbone	18.99	20.34	19.38	4.22	5.05	**12.16**	5.63	23.20	1.02
	SupCon	25.15	15.45	16.12	6.24	3.73	4.95	2.35	17.74	1.60
	Cross Entropy	11.83	6.91	7.65	2.08	1.92	0.64	1.47	7.18	1.00
	MS-loss	19.82	15.24	14.21	6.59	4.85	6.07	5.29	14.96	1.92
	MaskCon	**34.32**	21.75	18.03	13.16	9.05	8.63	5.00	**30.65**	10.06
	TaxoCon w/o EMA	22.49	14.23	16.94	6.25	4.14	4.63	3.82	16.72	1.86
	TaxoCon(SV)	23.96	21.75	**21.31**	7.44	5.22	8.31	**7.65**	18.33	2.41
	TaxoCon(MV)	31.66	**23.78**	19.67	**13.20**	**9.12**	8.15	6.76	30.35	**11.66**

Table 11: per-group linear-probing accuracy for ResNet-50

F.3 TAXONOMY STRUCTURE METRICS

Metric	Method	Mollusks	Mammals	Fishes	Insects	Birds	Reptiles	Amphibians	Fungi	Plants
Spearman	Frozen Backbone	18.21	30.84	19.73	23.18	17.42	26.15	29.66	20.07	20.84
	SupCon	16.33	28.25	13.64	22.16	14.17	14.57	27.28	17.46	20.66
	Cross Entropy	13.90	18.73	8.75	19.56	10.47	12.16	22.01	12.83	17.11
	MS-loss	16.33	26.77	15.26	21.49	14.59	20.53	26.88	17.78	20.71
	TaxoCon w/o EMA	16.12	28.69	14.26	21.95	14.16	13.66	27.11	17.54	22.62
	TaxoCon(SV)	17.86	31.11	20.42	**25.09**	17.55	27.44	30.31	19.37	20.31
	TaxoCon(MV)	**19.24**	**32.49**	**21.51**	24.92	**18.83**	**27.98**	**31.08**	**21.34**	**23.34**
Dendro-Tax	Frozen Backbone	16.94	26.81	13.68	22.43	15.72	24.81	26.34	18.67	16.13
	SupCon	15.78	24.70	12.64	21.88	12.97	14.20	26.75	16.75	**19.95**
	Cross Entropy	12.34	15.70	7.44	19.16	9.80	12.56	20.97	11.90	16.67
	MS-loss	14.40	21.41	13.95	21.04	12.59	18.82	25.52	15.28	15.66
	TaxoCon w/o EMA	15.18	25.59	12.91	23.01	12.29	13.03	26.26	16.42	19.46
	TaxoCon(SV)	16.29	**30.31**	17.95	**25.09**	16.56	25.36	27.88	16.26	19.11
	TaxoCon(MV)	**18.13**	29.00	**19.37**	24.25	**16.98**	**27.36**	**28.17**	**20.80**	19.89

Table 12: Per-group Spearman and Dendrogram–Taxonomy Agreement test results for ViT-S/16.

Metric	Method	Mollusks	Mammals	Fishes	Insects	Birds	Reptiles	Amphibians	Fungi	Plants
Spearman	Frozen Backbone	18.73	30.84	19.42	23.61	17.82	27.45	29.16	20.31	20.92
	SupCon	17.98	31.18	17.91	21.97	14.17	19.27	29.57	19.74	22.59
	Cross Entropy	15.32	26.02	12.93	22.20	10.47	15.45	30.32	16.55	21.26
	MS-loss	18.51	30.30	16.01	22.18	15.62	21.90	28.77	17.78	20.92
	TaxoCon w/o EMA	17.73	31.16	17.78	21.93	14.16	18.83	28.39	19.69	22.14
	TaxoCon(SV)	19.71	31.77	**22.04**	21.83	17.55	27.63	**30.73**	20.88	20.16
	TaxoCon(MV)	**19.92**	**32.05**	20.64	**24.94**	**18.95**	**29.12**	30.42	**21.53**	**24.23**
Dendro-Tax	Frozen Backbone	17.82	28.46	17.69	23.84	17.13	22.72	27.81	19.46	19.29
	SupCon	17.27	28.87	16.07	22.46	12.97	18.48	29.24	18.27	20.22
	Cross Entropy	12.87	23.49	11.54	21.74	9.80	14.50	29.92	15.79	18.12
	MS-loss	17.09	27.24	17.61	22.38	14.61	18.14	27.75	15.28	16.84
	TaxoCon w/o EMA	17.21	29.19	16.14	21.87	12.29	17.79	28.84	18.78	20.56
	TaxoCon(SV)	18.29	**31.65**	**20.72**	22.93	16.56	19.88	**31.86**	19.68	20.02
	TaxoCon(MV)	**18.76**	29.71	18.74	**25.06**	**18.16**	**26.28**	28.96	**20.17**	**22.02**

Table 13: Per-group Spearman and Dendrogram–Taxonomy Agreement test results for Swin-T.

Metric	Method	Mollusks	Mammals	Fishes	Insects	Birds	Reptiles	Amphibians	Fungi	Plants
Spearman	Frozen Backbone	15.43	26.27	14.83	19.74	13.91	15.06	27.58	16.87	9.88
	SupCon	16.67	26.67	14.94	20.99	14.05	16.41	29.17	17.23	11.07
	Cross Entropy	11.23	22.61	9.35	19.46	11.45	14.44	24.39	12.42	8.83
	MS-loss	14.06	25.51	10.14	21.05	12.37	16.34	26.70	14.01	6.72
	MaskCon	17.09	28.18	16.71	22.63	17.34	16.49	26.43	19.17	17.55
	TaxoCon w/o EMA	16.78	27.26	15.20	20.86	14.15	16.35	29.26	17.35	11.70
	TaxoCon(SV)	17.26	28.01	15.04	21.78	14.88	**17.61**	28.95	17.93	14.67
	TaxoCon(MV)	**18.61**	**30.22**	**18.06**	**22.86**	**18.10**	17.22	**29.80**	**19.92**	**19.63**
Dendro-Tax	Frozen Backbone	16.23	25.17	14.34	20.86	15.27	15.13	27.76	16.91	10.51
	SupCon	14.65	24.03	11.87	20.50	13.29	15.36	28.36	15.23	10.31
	Cross Entropy	8.80	19.90	7.76	19.37	11.84	14.77	23.69	11.62	8.58
	MS-loss	12.36	22.11	7.50	20.96	11.11	16.00	25.89	12.48	5.52
	MaskCon	13.44	22.89	13.40	20.90	14.73	15.91	25.94	16.38	12.01
	TaxoCon w/o EMA	15.04	24.91	13.19	20.13	13.25	**16.04**	28.32	15.35	10.15
	TaxoCon(SV)	14.57	24.92	13.20	20.49	13.64	14.34	27.59	16.05	13.13
	TaxoCon(MV)	**17.06**	**26.18**	**15.29**	**21.68**	**16.28**	15.57	**28.68**	**17.62**	**16.08**

Table 14: Per-group Spearman and Dendrogram–Taxonomy Agreement test results for ResNet-50.