

Unlabeled Echoes: Pseudo-Labels and Genus-Aware Smoothing for Bat Call Recognition

Frank Fundel¹ and Alexandra Howard²

¹ Ludwig-Maximilians-Universität München, Germany

² University of the Free State, Department of Zoology and Entomology, Qwaqwa Campus, South Africa

Abstract. Passive acoustic monitoring produces far more bat recordings than experts can label. We show that simple model-generated pseudo-labels turn this surplus into effective supervision. We compare pseudo-labeling with other semi-supervised learning methods on an 18-species European corpus using only 10% of its training labels, then transfer the strongest approaches to South African field audio containing nine bat taxa and a nuisance class. Pseudo-labeling outperforms the other semi-supervised learning methods on every European measure, recovering up to 61.5% of the gap to full supervision. It transfers to field audio with gains of 10.69 points in species accuracy and 4.96 points in species macro-F1. We also introduce genus-aware smoothing, which directs uncertain target mass toward congeneric species. Combined with uniform smoothing, it reaches 79.16 species macro-F1, 4.73 points above hard targets. Simple pseudo-labels are therefore highly effective at this ecological data scale, while genus-aware targets inject useful biological structure at no annotation cost. <https://code4conservation.github.io/UnlabeledEchoes/>.

Keywords: bioacoustics · bat-call recognition · semi-supervised learning · label smoothing · passive acoustic monitoring

1 Introduction

Bats provide pest control, pollination, and seed dispersal, while their sensitivity to environmental change makes them valuable indicators of ecosystem condition [15, 17]. Passive acoustic monitoring can sample entire nights of ultrasonic activity with little intervention. Yet a recording archive is not an ecological result: species-level labels still require scarce expertise, and annotation effort grows with every deployed recorder [8, 32]. Bat monitoring therefore presents a sharp asymmetry: collecting calls is increasingly cheap, but converting them into trustworthy supervision remains expensive.

Species recognition is intrinsically difficult. Bat echolocation is shaped by behavioral and environmental tasks, creating substantial within-species variation and between-species overlap; automated systems consequently identify acoustic groups more reliably than many individual species [29]. Congeneric species are

therefore biologically plausible alternatives, while devices, habitats, call phase, overlapping activity, and class imbalance further change the observed signal. Automatic recognition has progressed from hand-crafted call descriptors and shallow classifiers [24, 27, 35] to convolutional and Transformer models operating on time–frequency representations [5, 6, 19]. Modern models exploit labels more effectively but do not eliminate their cost; the largest resource is often the unannotated archive already being collected.

Semi-supervised methods are usually benchmarked on curated images or speech. Their ranking remains unclear for moderate-scale, multi-label ecological audio. Pseudo-labeling converts teacher predictions into task-specific targets and expands supervision *across recordings* [18]. Genus-aware smoothing reserves uncertain target mass for congeneric species and structures supervision *across classes*.

We test this claim in two complementary settings. The European study masks 90% of an 18-species training corpus; the South African study adds genuinely unlabeled field audio to labels for nine bat taxa and a nuisance output. This changes geography, recording conditions, and source vocabulary, testing whether the conclusion survives outside label masking. The classifier is based on the BioAcoustic Transformer (BAT) of Fundel et al. [6]; fixing the compact backbone isolates how supervision is constructed.

Our contributions are:

- We show that pseudo-labeling leads every measure in a controlled 10%-label comparison, recovering up to 61.5% of the gap to full supervision, and demonstrate that its species gains transfer to independent South African field audio.
- We introduce genus-aware smoothing, a target-level method that allocates uncertainty among congeneric species without additional annotations, and show in a controlled ablation that combining it with uniform smoothing improves species macro-F1 by 4.73 points over hard targets.

2 Related Work

Bat and ecological acoustics. Early identification used manually measured call features and shallow classifiers [24, 27]; deep systems now reduce feature engineering and scale to larger archives [5, 19]. Fundel et al. introduced BAT, the compact Transformer on which our classifier is based [6]. Most bat recognizers nevertheless remain supervised and inherit the biases of their annotations. Bat2Web evaluates an SGAN with reduced label fractions, although its additional samples are generator-produced rather than unlabeled field calls [21]. Recent work studies cross-region bat recognition and self-supervised bat-song representations [13, 20], while reviews continue to identify domain shift and label quality as central bottlenecks [32].

Bioacoustic benchmarks and pretrained models. BEANS, BirdSet, and iNatSounds provide cross-taxon, avian multi-label, and broad species-recognition benchmarks [3, 12, 26]. BirdNET, AVES, and the Perch family provide reusable

predictions or embeddings, including transfer from birds to bats and other taxa; Perch 2.0 and MetaPerch further broaden pretraining [2, 7, 11, 16, 22]. We instead fix an ultrasonic BAT backbone and vary how supervision is obtained.

Learning from unlabeled audio. SimCLR [4] and BYOL [10] learn invariant representations from paired augmentations; audio-native alternatives include BYOL-A, SSAST, and AVES [9, 11, 23]. They optimize a proxy representation before fitting the species layer; self-training instead creates supervision directly in that output space. Pseudo-labeling [18], Noisy Student [37], and FixMatch [31] differ mainly in teacher updates, perturbations, and consistency constraints. CAP and ML-FixMatch+DA add class-aware positive/negative confidence control and distribution alignment, components absent from the retained implementation [14, 36]. The families make different demands on augmentation, batches, and teacher reliability; we compare them at ecological scale and test whether their ranking transfers to field audio. Active learning instead spends expert effort on model-selected recordings [25]; it is complementary to our fixed-label setting, which asks how much can be learned without requesting new annotations.

Taxonomic supervision. Uniform label smoothing regularizes classifiers by moving probability mass away from a hard target [33]. Contextual smoothing and hierarchy-aware losses instead encode phylogenetic or semantic proximity [1, 34]. We bring this principle to multi-label bat-call recognition by concentrating smoothing mass on congeneric species, reflecting the documented ambiguity of species-level acoustic identification [29]. This target-level design requires neither an auxiliary hierarchy model nor further annotations, and remains compatible with ordinary uniform smoothing.

3 Method

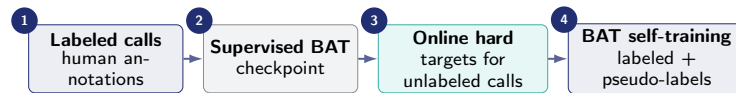


Fig. 1: Retained pseudo-labeling protocol. The supervised checkpoint generates online hard targets while training continues on interleaved labeled and unlabeled batches.

BAT backbone. Our classifier is based on the BioAcoustic Transformer (BAT) introduced by Fundel et al. [6]. BAT divides each spectrogram sequence into 60 overlapping patches and maps every patch to a 64-dimensional token with five convolution–batch-normalization–ReLU blocks. A learned classification token is appended before two self-attention layers with two heads and a feed-forward width of 32. A linear layer and sigmoid activation produce the multi-label species probabilities. For probability p_c and target y_c of class c , we use the asymmetric

loss [28]

$$\ell_{\text{ASL}}(\mathbf{p}, \mathbf{y}) = \sum_c [-y_c(1 - p_c) \log p_c - (1 - y_c)p_c^2 \log(1 - p_c)]. \quad (1)$$

The first term penalizes missed positive classes, while the second penalizes false positives; the factors $(1 - p_c)$ and p_c^2 reduce the influence of easy examples, especially abundant negatives. Optimizer and schedule details are given in Appendix A. The same compact backbone is retained across the comparison, isolating how supervision is constructed rather than changing model capacity.

Learning across recordings. A supervised BAT checkpoint is loaded and training continues on the same model; no separate frozen teacher is maintained (Fig. 1). At each self-training step, we set $\tilde{\mathbf{y}}$ to a one-hot vector: 1 for the highest-scoring class and 0 for every other class. Let $\hat{\mathbf{p}}$ be the current prediction. The unlabeled-data loss is scored with the ASL in Eq. 1:

$$\mathcal{L}_U = \lambda \ell_{\text{ASL}}(\hat{\mathbf{p}}, \tilde{\mathbf{y}}). \quad (2)$$

Here λ controls the contribution of the unlabeled example; it is initialized at zero and increased to one during training. Targets are regenerated online, not stored, and no confidence threshold is used. Human-labeled batches use the same ASL without the unlabeled-data weight. Validation and test data are kept separate for evaluation.

Learning across classes. Uniform smoothing treats every wrong class alike. Our *genus-aware smoothing* instead places uncertain mass only on species that share the target genus (Fig. 2a). For a multi-hot label $\mathbf{y}$, let m_k count how often class k is a congeneric alternative and let r_k normalize these counts:

$$m_k = \sum_{c: y_c=1} \mathbb{I}[k \in G(c) \setminus \{c\}], \quad r_k = \frac{m_k}{\max(\sum_j m_j, 1)}, \quad (3)$$

$$q_k = (1 - \beta)y_k + \beta\|\mathbf{y}\|_1 r_k.$$

Thus β moves target mass to related species rather than distributing it uniformly. The retained combined recipe uses genus smoothing with $\beta = .10$, followed by ordinary uniform smoothing with $\alpha = .02$.

4 Experiments

4.1 Task and Data

Given a time–frequency representation x , the model predicts probabilities $\hat{\mathbf{p}} \in [0, 1]^C$ for a multi-hot target $\mathbf{y} \in \{0, 1\}^C$. We evaluate *single-species* sequences with accuracy and unweighted per-species F1, and *synthetic mixtures* made by combining one to three sequences and taking the union of their labels. Mixtures use macro-F1 and test multi-label interference rather than natural acoustic overlap.

Skiba masks labels within a curated collection, enabling comparison with full supervision; UFS adds independent 256-kHz field audio whose 17-species source archive may contain taxa outside the model’s ten outputs. Both use the same spectrogram task and compact model but test different kinds of unlabeled data; exact counts are in Table 1.

Table 1: Datasets used in the experiments. Splits are made at the recording level before sequence extraction with random seed 42.

Data	Source archive	Model outputs	Sequence split	Supervision used
Skiba	>1,500 recordings; 29 species; 96 kHz [30]	18 species	19,030 total: 11,186 train, 2,890 validation, 4,954 test	1,119 labeled train (10%); 10,067 train (90%) treated as unlabeled; full reference uses all 11,186
UFS	>40k recordings; 17 species; ~15k without labels; 256 kHz	9 bat taxa + Pesticide Spray	15,863 total: 6,736 train, 3,857 validation, 5,270 test	30,000 additional unlabeled field sequences

4.2 Experimental Protocol

Input construction. Recordings are assigned to the training, validation, and test sets before sequences are created, so sequences from the same recording cannot appear in different sets. Ultrasonic audio is time-dilated, high-pass filtered, downsampled, and converted to a decibel spectrogram. Training mixtures combine one to three sequences and union their labels; labeled and unlabeled examples use the same acoustic construction.

Splits and supervision. The preparation script splits recordings within each metadata class into 60% training, 20% validation, and 20% test sets (seed 42) before extracting and shuffling sequences. The European comparison uses 10% human labels and treats the other 90% as unlabeled; the 100% model is an upper reference. BYOL, pseudo-labeling, and FixMatch are then evaluated on UFS.

Evaluation and reporting. Species F1 is the unweighted mean of the per-species F1 scores. The BYOL row reports a four-run mean; remaining results are point estimates, and all gains below are absolute percentage points.

Comparison methods. Autoencoder pretraining reconstructs spectrogram patches, while BYOL aligns augmented views before classifier fitting [10]. Noisy Student uses a perturbed teacher and student [37]; FixMatch uses weak-strong consistency [31]. We also evaluate both cascade orders. These are comparison

baselines; pseudo-labeling and genus-aware smoothing are the methods proposed here.

4.3 Results

Table 2: Skiba results (%). Except for the gray full-supervision reference, all rows use 10% labels; ✓ denotes unlabeled calls. Bold marks the best low-label result.

Method	Labeled	Unlab.	Mixture F1 ↑	Species acc. ↑	Species F1 ↑
Full supervision	100%	✗	72.46	84.00	80.33
10% baseline	10%	✗	58.74	75.06	62.47
Autoencoder	10%	✓	44.74	68.53	53.61
BYOL	10%	✓	52.68	69.54	57.69
Noisy Student	10%	✓	59.71	75.22	67.49
FixMatch	10%	✓	61.72	76.06	68.71
Pseudo-label → FixMatch	10%	✓	63.39	76.58	69.80
FixMatch → Pseudo-label	10%	✓	62.18	76.66	67.97
Pseudo-label	10%	✓	65.12	80.48	73.45

Pseudo-labels recover up to 61.5% of lost supervision. Pseudo-labeling leads all three European measures (Table 2), adding 6.38 mixture-F1, 5.42 accuracy, and 10.98 species-F1 points over the matched baseline. This closes **46.5%**, **60.6%**, and **61.5%** of the respective gaps to full supervision.

Table 3: UFS results (%). All methods use 6,736 labeled sequences; ✓ adds 30,000 unlabeled field sequences. Bold marks the best result.

Method	Unlab.	Mixture F1 ↑	Species acc. ↑	Species F1 ↑
Supervised baseline	✗	55.25	67.55	48.65
BYOL	✓	55.82	75.29	49.01
FixMatch	✓	57.98	76.99	49.04
Pseudo-label	✓	56.69	78.24	53.61

Pseudo-labeling transfers to field audio. Every transferred learner exceeds the UFS baseline on all measures (Table 3). Pseudo-labeling gives the largest species gains (**+10.69 accuracy**, **+4.96 macro-F1**), while FixMatch leads on mixtures (+2.73 macro-F1).

Controlled smoothing ablation. All four runs in Fig. 2 use the same data split, compact backbone, preprocessing, optimizer, and training schedule; only the construction of the training targets differs. Genus-aware smoothing alone

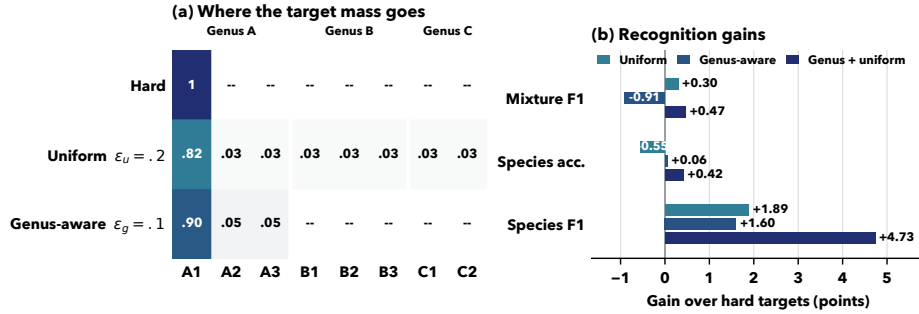


Fig. 2: Genus-aware smoothing and controlled compact-backbone results. (a) Target mass is restricted to related species. (b) Genus plus uniform smoothing gives the largest gains.

raises species macro-F1 from 74.43 to 76.03 (+1.60 points). The combined recipe is strongest on every measure: 67.79 mixture F1 (+0.47), 81.51 species accuracy (+0.42), and **79.16 species F1 (+4.73)**; it also exceeds uniform smoothing by 2.84 species-F1 points. Because accuracy weights sequences while macro-F1 weights species equally, the +0.42 versus +4.73 pattern indicates that smoothing chiefly benefits weaker classes.

5 Discussion

Across both experiments, pseudo-labeling is the most consistent way to exploit additional audio with the fixed compact BAT backbone. It is the strongest low-label method on Skiba, recovering up to 61.5% of the full-supervision gap, and improves the independent UFS field pool by 10.69 accuracy and 4.96 species-macro-F1 points. This transfer spans geography and recording conditions, while UFS remains limited to the model’s known output classes.

Genus-aware smoothing complements pseudo-labeling by assigning uncertainty to biologically plausible congeneric alternatives. With the split, backbone, preprocessing, optimizer, and schedule fixed, its 4.73-point macro-F1 gain isolates target construction; the smaller 0.42-point accuracy gain indicates better performance on weaker species. These results use one European corpus, one field pool, and a fixed compact backbone, with hard online labels and a closed output vocabulary. Deployment should therefore pair predictions with confidence or open-set checks and expert review.

6 Conclusion

This study shows that bat archives support semi-supervised recognition without exhaustive labeling. With 10% of the European labels, pseudo-labeling recovers up to 61.5% of the full-supervision gap and transfers to independent South African field audio. Genus-aware smoothing adds biologically structured uncertainty and improves species macro-F1 by 4.73 points without additional annotations. Together, these methods make unlabeled archives a practical resource for bat monitoring, with expert review for unfamiliar calls.

References

1. Bertinetto, L., Mueller, R., Tertikas, K., Samangoeei, S., Lord, N.A.: Making better mistakes: Leveraging class hierarchies with deep networks. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 12506–12515 (2020). <https://doi.org/10.1109/CVPR42600.2020.01255>
2. Chasmai, M., Dumoulin, V., Hamer, J.: MetaPerch: Learning from metadata for bioacoustics foundation models. In: Proceedings of the 43rd International Conference on Machine Learning (2026)
3. Chasmai, M., Shepard, A., Maji, S., Van Horn, G.: The iNaturalist sounds dataset. In: Advances in Neural Information Processing Systems. vol. 37 (2024). <https://doi.org/10.52202/079017-4213>
4. Chen, T., Kornblith, S., Norouzi, M., Hinton, G.: A simple framework for contrastive learning of visual representations. In: Proceedings of the 37th International Conference on Machine Learning. Proceedings of Machine Learning Research, vol. 119, pp. 1597–1607. PMLR (2020)
5. Chen, X., Zhao, J., Chen, Y.h., Zhou, W., Hughes, A.C.: Automatic standardized processing and identification of tropical bat calls using deep learning approaches. *Biological Conservation* **241**, 108269 (2020). <https://doi.org/10.1016/j.biocon.2019.108269>
6. Fundel, F., Braun, D.A., Gottwald, S.: Automatic bat call classification using transformer networks. *Ecological Informatics* **78**, 102288 (2023). <https://doi.org/10.1016/j.ecoinf.2023.102288>
7. Ghani, B., Denton, T., Kahl, S., Klinck, H.: Global birdsong embeddings enable superior transfer learning for bioacoustic classification. *Scientific Reports* **13**, 22876 (2023). <https://doi.org/10.1038/s41598-023-49989-z>
8. Gibb, R., Browning, E., Glover-Kapfer, P., Jones, K.E.: Emerging opportunities and challenges for passive acoustics in ecological assessment and monitoring. *Methods in Ecology and Evolution* **10**(2), 169–185 (2019). <https://doi.org/10.1111/2041-210X.13101>
9. Gong, Y., Lai, C.I., Chung, Y.A., Glass, J.: SSAST: Self-supervised audio spectrogram transformer. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 36, pp. 10699–10709 (2022). <https://doi.org/10.1609/aaai.v36i10.21315>
10. Grill, J.B., Strub, F., Altché, F., Tallec, C., Richemond, P.H., Buchatskaya, E., Doersch, C., Avila Pires, B., Guo, Z.D., Azar, M.G., Piot, B., Kavukcuoglu, K., Munos, R., Valko, M.: Bootstrap your own latent: A new approach to self-supervised learning. In: Advances in Neural Information Processing Systems. vol. 33, pp. 21271–21284 (2020)
11. Hagiwara, M.: AVES: Animal vocalization encoder based on self-supervision. In: ICASSP 2023—2023 IEEE International Conference on Acoustics, Speech and Signal Processing. pp. 1–5. IEEE (2023). <https://doi.org/10.1109/ICASSP49357.2023.10095642>
12. Hagiwara, M., Hoffman, B., Liu, J.Y., Cusimano, M., Effenberger, F., Zacarian, K.: BEANS: The benchmark of animal sounds. In: ICASSP 2023—2023 IEEE International Conference on Acoustics, Speech and Signal Processing. pp. 1–5. IEEE (2023). <https://doi.org/10.1109/ICASSP49357.2023.10096686>
13. de Heer Kloots, M., Knörnschild, M.: Exploring bat song syllable representations in self-supervised audio encoders. arXiv preprint arXiv:2409.12634 (2024).

- <https://doi.org/10.48550/arXiv.2409.12634>, presented at the 4th International Workshop on Vocal Interactivity in-and-between Humans, Animals and Robots
14. Ihler, S., Kuhnke, F., Kuhlitz, T., Seel, T.: Distribution-aware multi-label FixMatch for semi-supervised learning on CheXpert. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops. pp. 2295–2304 (2024). <https://doi.org/10.1109/CVPRW63382.2024.00235>
 15. Jones, G., Jacobs, D.S., Kunz, T.H., Willig, M.R., Racey, P.A.: Carpe noctem: The importance of bats as bioindicators. *Endangered Species Research* **8**, 93–115 (2009). <https://doi.org/10.3354/esr00182>
 16. Kahl, S., Wood, C.M., Eibl, M., Klinck, H.: BirdNET: A deep learning solution for avian diversity monitoring. *Ecological Informatics* **61**, 101236 (2021). <https://doi.org/10.1016/j.ecoinf.2021.101236>
 17. Kunz, T.H., de Torrez, E.B., Bauer, D., Lobova, T., Fleming, T.H.: Ecosystem services provided by bats. *Annals of the New York Academy of Sciences* **1223**(1), 1–38 (2011). <https://doi.org/10.1111/j.1749-6632.2011.06004.x>
 18. Lee, D.H.: Pseudo-label: The simple and efficient semi-supervised learning method for deep neural networks. In: ICML Workshop on Challenges in Representation Learning (2013)
 19. Mac Aodha, O., Gibb, R., Barlow, K.E., Browning, E., Firman, M., Freeman, R., Harder, B., Kinsey, L., Mead, G.R., Newson, S.E., Pandourski, I., Parsons, S., Russ, J., Szodoray-Paradi, A., Szodoray-Paradi, F., Tilova, E., Girolami, M., Brostow, G., Jones, K.E.: Bat detective—deep learning tools for bat acoustic signal detection. *PLOS Computational Biology* **14**(3), e1005995 (2018). <https://doi.org/10.1371/journal.pcbi.1005995>
 20. Mac Aodha, O., Martínez Balvanera, S., Damstra, E., Cooke, M., Eichinski, P., Browning, E., Gilmour, L., Staniewicz, A., Barataud, M., Boughey, K., Coles, R., Giacomini, G., Mac Swiney G., M.C., Obrist, M.K., Parsons, S., Sattler, T., Jones, K.E.: Towards a general approach for bat echolocation detection and classification. *bioRxiv* (2026). <https://doi.org/10.1101/2022.12.14.520490>, preprint, version 2, posted 16 May 2026
 21. Mahbub, T., Bhagwagar, A., Chand, P., Zualkernan, I., Judas, J., Dghaym, D.: Bat2Web: A framework for real-time classification of bat species echolocation signals using audio sensor data. *Sensors* **24**(9), 2899 (2024). <https://doi.org/10.3390/s24092899>
 22. van Merriënboer, B., Dumoulin, V., Hamer, J., Harrell, L., Burns, A., Denton, T.: Perch 2.0: The bitter lesson for bioacoustics. *arXiv preprint arXiv:2508.04665* (2025). <https://doi.org/10.48550/arXiv.2508.04665>
 23. Niizumi, D., Takeuchi, D., Ohishi, Y., Harada, N., Kashino, K.: BYOL for audio: Self-supervised learning for general-purpose audio representation. In: 2021 International Joint Conference on Neural Networks. pp. 1–8. IEEE (2021). <https://doi.org/10.1109/IJCNN52387.2021.9534474>
 24. Parsons, S., Jones, G.: Acoustic identification of twelve species of echolocating bat by discriminant function analysis and artificial neural networks. *Journal of Experimental Biology* **203**(17), 2641–2656 (2000). <https://doi.org/10.1242/jeb.203.17.2641>
 25. Qian, K., Zhang, Z., Baird, A., Schuller, B.: Active learning for bird sounds classification. *Acta Acustica united with Acustica* **103**(3), 361–364 (2017). <https://doi.org/10.3813/AAA.919064>

26. Rauch, L., Schwinger, R., Wirth, M., Heinrich, R., Huseljic, D., Herde, M., Lange, J., Kahl, S., Sick, B., Tomforde, S., Scholz, C.: BirdSet: A large-scale dataset for audio classification in avian bioacoustics. In: International Conference on Learning Representations (2025)
27. Redgwell, R.D., Szewczak, J.M., Jones, G., Parsons, S.: Classification of echolocation calls from 14 species of bat by support vector machines and ensembles of neural networks. *Algorithms* **2**(3), 907–924 (2009). <https://doi.org/10.3390/a2030907>
28. Ridnik, T., Ben-Baruch, E., Zamir, N., Noy, A., Friedman, I., Protter, M., Zelnik-Manor, L.: Asymmetric loss for multi-label classification. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 82–91 (2021). <https://doi.org/10.1109/ICCV48922.2021.00015>
29. Rydell, J., Nyman, S., Eklöf, J., Jones, G., Russo, D.: Testing the performances of automated identification of bat echolocation calls: A request for prudence. *Ecological Indicators* **78**, 416–420 (2017). <https://doi.org/10.1016/j.ecolind.2017.03.023>
30. Skiba, R.: Europäische Fledermäuse: Kennzeichen, Echoortung und Detektoranwendung. Die Neue Brehm-Bücherei, Westarp Wissenschaften, Hohenwarsleben (2003)
31. Sohn, K., Berthelot, D., Li, C.L., Zhang, Z., Carlini, N., Cubuk, E.D., Kurakin, A., Zhang, H., Raffel, C.: FixMatch: Simplifying semi-supervised learning with consistency and confidence. In: Advances in Neural Information Processing Systems. vol. 33, pp. 596–608 (2020)
32. Stowell, D.: Computational bioacoustics with deep learning: A review and roadmap. *PeerJ* **10**, e13152 (2022). <https://doi.org/10.7717/peerj.13152>
33. Szegedy, C., Vanhoucke, V., Ioffe, S., Shlens, J., Wojna, Z.: Rethinking the inception architecture for computer vision. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 2818–2826 (2016). <https://doi.org/10.1109/CVPR.2016.308>
34. Trammell, M.J., Oberoi, P., Egenrieder, J.A., Kaufhold, J.: Contextual label smoothing with a phylogenetic tree on the iNaturalist 2018 challenge dataset. *Journal of the Washington Academy of Sciences* **105**(1), 23–46 (2019)
35. Walters, C.L., Freeman, R., Collen, A., Dietz, C., Fenton, M.B., Jones, G., Obrist, M.K., Puechmaille, S.J., Sattler, T., Siemers, B.M., Parsons, S., Jones, K.E.: A continental-scale tool for acoustic identification of european bats. *Journal of Applied Ecology* **49**(5), 1064–1074 (2012). <https://doi.org/10.1111/j.1365-2664.2012.02182.x>
36. Xie, M.K., Xiao, J., Liu, H.Z., Niu, G., Sugiyama, M., Huang, S.J.: Class-distribution-aware pseudo-labeling for semi-supervised multi-label learning. In: Advances in Neural Information Processing Systems. vol. 36, pp. 25731–25747 (2023). <https://doi.org/10.52202/075280-1119>
37. Xie, Q., Luong, M.T., Hovy, E., Le, Q.V.: Self-training with noisy student improves ImageNet classification. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 10687–10698 (2020). <https://doi.org/10.1109/CVPR42600.2020.01070>

A Implementation details

The retained pseudo-label implementation interleaves unlabeled and labeled updates. A full pass over the human-labeled loader follows every ten unlabeled mini-batches. The pseudo-label weight is $\lambda_t = 0$ before step 23, increases linearly as $(t - 23)/10$ from steps 23 to 33, and remains 1 thereafter; the retained run starts at the beginning of this ramp. Human-labeled updates use unweighted asymmetric loss. The recording-level split and sequence-level masking are described in the main protocol and Table 1.

All main runs use SGD with a 5×10^{-4} learning rate, cosine decay, and 46 epochs. The BYOL row is a four-run mean, whereas the remaining rows are point estimates.