Bridging Domain Gaps for Fine-Grained Moth Classification Through Expert-Informed Adaptation and Foundation Model Priors

Ross J. Gardiner[1] Guillaume Mougeot[2] Sareh Rowlands[1]
Benno I. Simmons[1] Flemming Helsing[2] Toke Thomas Høye[2]
[1]University of Exeter [2]Aarhus University

Abstract

Labelling images of Lepidoptera (moths) from automated camera systems is vital for understanding insect declines. However, accurate species identification is challenging due to domain shifts between curated images and noisy field imagery. We propose a lightweight classification approach, combining limited expert-labelled field data with knowledge distillation from the high-performance BioCLIP2 foundation model into a ConvNeXt-tiny architecture. Experiments on 101 Danish moth species from AMI camera systems demonstrate that BioCLIP2 substantially outperforms other methods and that our distilled lightweight model achieves comparable accuracy with significantly reduced computational cost. These insights offer practical guidelines for the development of efficient insect monitoring systems and bridging domain gaps for fine-grained classification.

1. Introduction

Insects are vital to terrestrial ecosystems and global agriculture [18]. They constitute roughly half of all animal species [29], and terrestrial arthropods collectively account for 20 times the biomass of all wild mammals and birds combined [2]. Despite their importance, insect populations are declining worldwide [31, 33], this remains poorly understood due to a lack of scalable field monitoring methods [20].

In response, automated insect camera traps [4, 10, 27] autonomously photograph insects in-situ, enabling long-term and resource-efficient population monitoring [26]. But, the resulting raw imagery must be processed before it can provide actionable ecological insights. Advances in detection [5, 6] and classification [7, 23] have made computer vision the leading approach for this task [13].

Species-level recognition from camera-trap images remains difficult: morphological differences are subtle and expert taxonomists are scarce [9], making annotation expensive. Citizen science repositories such as the Global Biodiversity Information Facility (GBIF) [1] provide labelled images for many taxa and are popular sources of training data. However, domain shift (differences in pose, lighting, and image quality between GBIF and automated camera trap imagery) reduces performance when models are trained exclusively on "source-domain" GBIF images and evaluated on "target-domain" in-situ insects [14].

In addition, developing *lightweight* computer vision models for insect monitoring remains important because: (1) ecologists worldwide may lack the computing infrastructure to run large models efficiently; (2) lightweight models are easier to retrain, which is valuable as most insect species remain undiscovered [29] and new species must be integrated; (3) lightweight architectures facilitate 'edge' computing directly on camera trap hardware in remote environments. This is common in insect camera trap designs [3, 8, 26] where on-device inference is critical when connectivity constraints prohibit remote inference. Edge inference also allows the detection of specific species of interest on the device, allowing for the removal of unwanted images, conserving storage space for long deployments [10, 25].

We develop and evaluate a lightweight architecture that addresses domain shift through two strategies: (1) integrating expert-labelled target-domain data into training, and (2) guiding learning via knowledge distillation (KD) from large-scale pretrained foundation models. We evaluate on a fine-grained and domain-shifted dataset of moth images from Automated Monitoring of Insects (AMI) cameras [14] deployed around 12 sites in Denmark. Combining expert-labelled target-domain data and KD enables our lightweight model to achieve comparable performance to BioCLIP and BioCLIP2, despite significantly lower parameter count. We provide recommendations for ecologists to use foundation models for their applications, and for the computer vision community towards fine-grained domain adaptation.

1.1. Related Work

There are many established methods for domain adaptation [34], including feature alignment, adversarial learning and self-training. Synthetic data generation [15] has also previously been shown to improve generalisation for insect

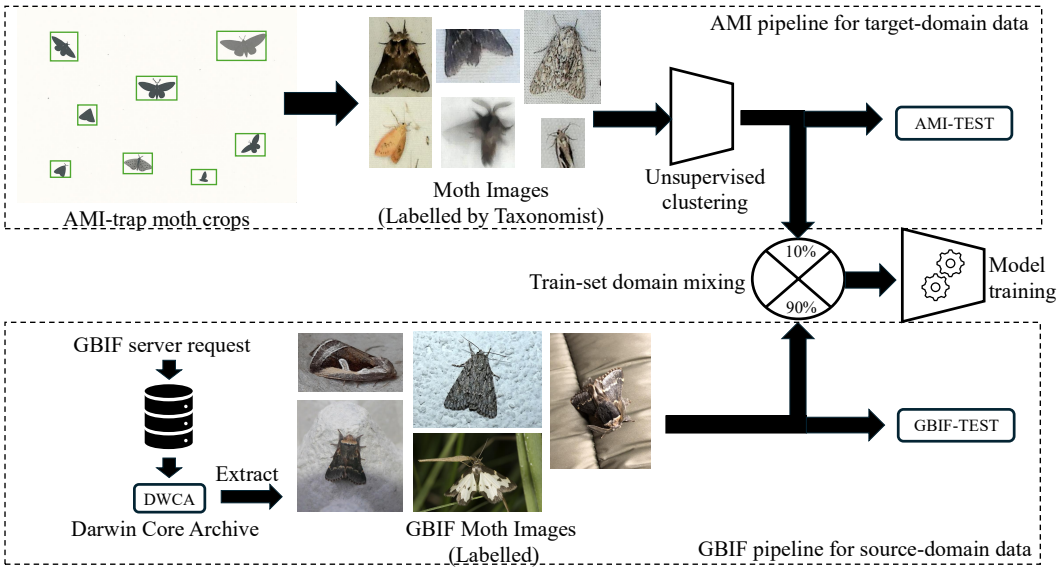

Figure 1. Data processing pipeline showing how moth images from AMI (target-domain) and GBIF (source-domain) are curated, labelled at the species level, and split into training and test sets. Controlled mixing of domains enables analysis of domain shift effects. Example images illustrate differences between the two domains.

images. Here, we explore domain adaptation through supervised mixing of the training set as it is straightforward to implement and understand and allows us to leverage the limited expert-labelled target data available in our scenario.

The use of CLIP-style [24] foundation models offers another route to cross-domain generalisation, pre-training large-scale representations transferable to diverse target domains. For biological images, recent models such as Insect-Foundation [22], CLIBD [11] and Biotrove [35] have shown strong performance on fine-grained insect identification tasks. Notably, BioCLIP [28], trained on the TreeOfLife-10M dataset, and its successor BioCLIP2 [12], scaled to 214 million mostly GBIF-sourced images, achieve state-of-the-art accuracy. To retain the broad domain knowledge of such models while adapting to specific target tasks, we use KD, training smaller student model is to match the features of a larger pre-trained teacher [21]. In this context, KD serves as a domain adaptation strategy, transferring generalisable representations from teacher to student [16].

Our work is related to [14], which introduced the AMI dataset and demonstrated ConvNeXt [17] performance on cross-domain insect classification. We benchmark foundation models using a balanced dataset of Danish moth species and explore specialist model training with limited domain-specific data. Foundation models have also been used for camera trap classification; [32] applied pre-trained models for zero-shot vertebrate identification, showing strong adaptability. We extend this by evaluating BioCLIP2 on AMI insect images, assessing the impact of expert-labelled target data and knowledge distillation for smaller models.

2. Dataset

We used a dataset of images of moths species collected from 12 AMI systems deployed in three regions of Denmark across three years (2022-2024) [3]. Images captured by AMI are high resolution photographs of a large white screen, where moths are attracted to a UV light. Moths themselves make up small portions of each image and pre-existing object detection methods give crops for each moth instance [20]. Crops can be highly variable in size, lighting and quality, Figure 1 shows some examples. And, following the long-tailed distribution typical of ecological scenarios, there are typically many images of the most common species, and relatively few of the others [30]. Identifiable moths in the AMI images have been labelled to the species level by one of the authors, an expert insect taxonomist specialised in moth species identification. As a primary focus in this study is domain shift, for our experiments, we use a subset of AMI moth crops represented by 101 species, where each species is represented by 110 AMI images, to isolate and study domain-shift effects. We refer to these throughout as the "target-domain" examples.

Individual images in the target-domain can be highly correlated, as data are obtained by time-lapse, and frequently an individual is stationary for long periods and is therefore imaged repeatedly in the same position. To prevent leakage of highly correlated images from the training set into our test set, we use an unsupervised deep clustering method to split the train/test set for each class (see Supplementary Section 8). Of the 110 images per class, a test set of 10 is withheld, leaving 100 per class for training.

To leverage GBIF images, we download each directly using an automated toolkit, supplying the GBIF species ID for each AMI class and filtering for the 'imago' or adult life-stage, matching AMI life-stage. For each class, we obtained 224 images from unique GBIF occurrences, forming the "source-domain" dataset. These are randomly split into 184 training and 20 withheld test images per class.

We progressively increase the proportion of target-domain images in training. At each step, we include as many GBIF images as possible to maximise training size. Once all available target images are used (over 35%), further increases in the target proportion require reducing GBIF images. We constructed eight training sets with target-domain contributions of 0%, 1%, 5%, 10%, 20%, 25%, 33%, and 50% (statistics in Supplementary Section 9).

3. Methodology

We train four models for moth species classification across domain-mixed datasets. To benchmark BioCLIP and BioCLIP2, we train linear classifiers on their frozen vision encoder outputs (as shown in Figure 2). We also train a partially fine-tuned ConvNeXt-tiny model, unfreezing its classification head and top two layers, initialised with ImageNet-1K weights; this setup was found to perform best empirically. Finally, we apply KD from BioCLIP2 to ConvNeXt-tiny to improve generalisation through richer feature representations.

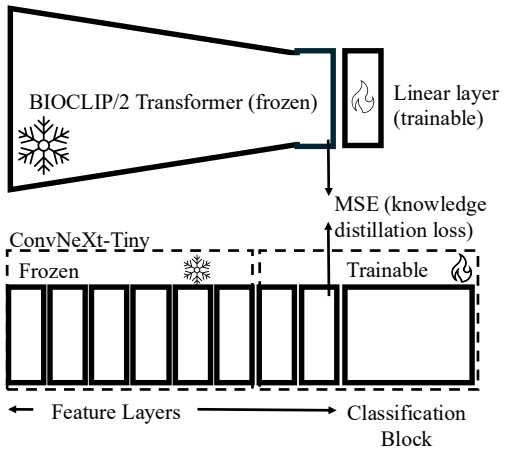

Figure 2. Foundation vision encoders (top) with a trainable classification linear layer. ConvNeXt instance (bottom) with pre-trained feature layers 1-6 frozen, the final feature layer is used for the mean-squared error (MSE) distillation loss for ConvNeXt+KD.

3.1. Knowledge Distillation

We use feature-based KD [21], where the student network aims to match the learned feature representations from the teacher to its own internal embeddings. Specifically, we implement a 'hint' loss, defined as the mean squared error between the final feature representation of ConvNeXt and the output embedding of BioCLIP2, defined by Equation (1):

$$\mathcal{L}_{\text{hint}} = \frac{1}{N}\sum_{i=1}^{N}(s_i - t_i)^2 \tag{1}$$

Where $N = B \times C$, B is the total number of elements in the batch and C are the dimensions of the features, s and t represent the student and teacher embeddings, respectively.

The hint loss is integrated into the ConvNeXt model loss using a weighting, defined by α, to determine the contribution from categorical cross-entropy loss from the model output and the hint loss from features supervised by the teacher. Our loss function is described by Equation (2):

$$\mathcal{L}_{\text{total}} = \alpha \cdot \mathcal{L}_{\text{CE}} + (1 - \alpha) \cdot \mathcal{L}_{\text{hint}} \tag{2}$$

Where $\mathcal{L}_{\text{CE}}$ is the categorical cross-entropy loss. In our experiments, we set α equal to 0.5, assuming that classification and hint learning are equally important for our task.

3.2. Training

All non-KD models are trained using categorical cross-entropy loss only. Each model is trained using an NVIDIA A100 GPU with the following hyperparameters: the learning rate, μ, is set to $1e-3$, a weight decay of $1e-5$ is applied, the mini-batch size is 64 and the AdamW optimiser is used [19]. Following [14], we also deploy a MixRes strategy to augment the GBIF source-domain images only. For MixRes, given an image, it has a 25% chance to be down-scaled to a size of 75x75 pixels or a 25% chance to be downscaled to 150x150 pixels. Further augmentations are applied to all images though PyTorch `RandAugment` function, with the `num_ops` variable set to 2 and the `magnitude` set to 3, each image is also 50% likely to be horizontally flipped. All images are then resized to the model input size (224x224).

4. Results

For each dataset and architecture, we train for 10 epochs and report top-1 classification accuracy on the target and source test sets in Table 1. Figure 3 shows target-domain accuracy; source-domain accuracy is discussed separately below.

BioCLIP2 consistently achieves the highest accuracy across all target-mix levels, particularly under low target supervision, with an average gain of 2.1% over the range 1–50% target data. ConvNeXt, ConvNeXt+KD, and BioCLIP show larger relative improvements with additional target data (average gains of 18.8%, 16.8%, and 9.1%, respectively). ConvNeXt and ConvNeXt+KD perform poorly with minimal target supervision but match or exceed BioCLIP with more target supervision. Knowledge distillation

	ConvNeXt-tiny (28M Params.)		ConvNeXt-tiny+KD (28M Params.)		BioCLIP (86M Params.)		BioCLIP2 (304M Params.)	
Target-domain Mix (%)	Top-1 acc. (target) (%)	Top-1 acc. (source) (%)	Top-1 acc. (target) (%)	Top-1 acc. (source) (%)	Top-1 acc. (target) (%)	Top-1 acc. (source) (%)	Top-1 acc. (target) (%)	Top-1 acc. (source) (%)
0%	59.4	88.1	64.7	91.2	71.2	95.2	88.3	97.6
1%	60.4	86.9	63.0	90.4	74.4	95.1	87.4	98.3
5%	72.8	87.1	79.4	90.9	76.6	95.0	89.4	98.4
10%	77.7	87.6	81.0	90.6	78.0	95.1	91.5	98.5
20%	83.0	87.9	85.0	90.8	81.1	95.0	90.0	98.0
25%	82.5	87.1	86.6	91.0	82.3	94.9	91.5	98.0
33%	85.9	88.1	86.4	89.9	83.9	95.1	91.3	98.0
50%	85.4	84.5	89.4	88.8	85.8	94.4	91.6	97.8

Table 1. Showing top 1 source and target accuracy (acc.) percentages for each model architecture at each level of target-domain supervision. Model parameter counts (params) in millions (M) are also shown on the top row, beneath architecture names.

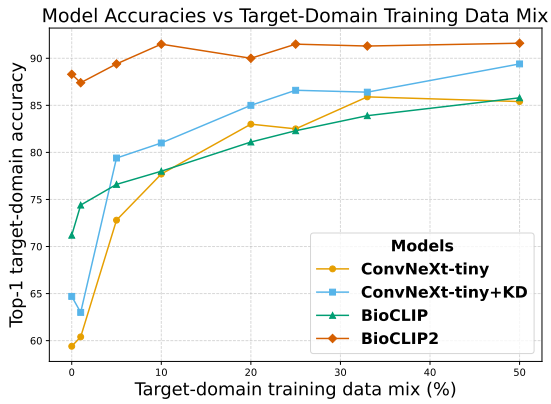

Figure 3. Top-1 target-domain accuracy plotted over training target-domain mix percentages for each architecture.

provides a consistent boost across all target-mix levels, averaging +3.6% over ConvNeXt.

Across most settings, source-domain accuracy remains higher than target for all models. For BioCLIP2, source accuracy is largely stable across target-mix levels. For the ConvNeXt variants and BioCLIP, there is a slight downward trend as more target data is included, reflecting the reduced amount of source-domain data in the mixed training set. The largest drop occurs at the 50% mix. At this point, ConvNeXt and ConvNeXt+KD are the only cases where target accuracy marginally surpasses source accuracy.

5. Discussion

BioCLIP2 consistently achieves the highest performance, underscoring its strong generalisation capabilities and robustness to domain shift. This can be attributed to its pre-training on a very large and diverse image corpus, which likely exposed the model to a wide range of morphological traits, lighting conditions, and poses. Therefore, we recommend BioCLIP2 as a model with high transferability to AMI images. BioCLIP, while weaker overall, performs surprisingly well in low target-supervision settings, suggesting that its pre-trained features are also highly transferable to our target domain, but less so than BioCLIP2.

Even modest target supervision of 5% significantly boosts target accuracy across all models, with particularly large gains of 13.4% for ConvNeXt and 14.7% for ConvNeXt+KD. The addition of KD from BioCLIP2 further enhances performance, suggesting it helps ConvNeXt learn deeper, target-relevant features beyond what target supervision provides. ConvNeXt+KD achieved the best accuracy per parameter, making it more suitable for compute-limited settings. Notably, at 50% target supervision, it matched BioCLIP2's performance without target supervision, despite having 10 times fewer parameters (28M vs. 304M).

Overall, we provide two insights: (1) when the requirement for lightweight models is most important, we recommend KD and train-time domain mixing as effective strategies to build performant models; (2) in scenarios where ample computing power is available and/or target domain data are especially limited, we advocate using BioCLIP2 and saving precious expert-labelled data for model evaluation.

6. Conclusion

Through our experiments, we arrive at new directions for in-situ moth classification under domain shift. We show foundation models as accurate, adaptable classifiers and develop methods to exploit BioCLIP2 for training lightweight models. These advances are critical for the development of insect camera traps, which are a vital tool for understanding global insect declines.

7. Acknowledgements

This article is based upon work from a short-term scientific mission (STSM) through the COST Action InsectAI, CA22129, supported by COST (European Cooperation in Science and Technology). This work was also funded via a doctoral training grant awarded as part of the UKRI AI Centre for Doctoral Training in Environmental Intelligence (UKRI grant number EP/S022074/1).

References

[1] Gbif home page, 2025. 1

[2] Yinon M. Bar-On, Rob Phillips, and Ron Milo. The biomass distribution on earth. *Proceedings of the National Academy of Sciences*, 115(25):6506–6511, 2018. 1

[3] Kim Bjerge, Henrik Karstoft, and Toke T. Høye. Towards edge processing of images from insect camera traps. *Remote Sensing in Ecology and Conservation*. 1, 2

[4] Kim Bjerge, Jakob Bonde Nielsen, Martin Videbæk Sepstrup, Flemming Helsing-Nielsen, and Toke Thomas Høye. An automated light trap to monitor moths (lepidoptera) using computer vision-based tracking and deep learning. *Sensors*, 21(2), 2021. 1

[5] Kim Bjerge, Jamie Alison, Mads Dyrmann, Carsten Eie Frigaard, Hjalte M. R. Mann, and Toke Thomas Høye. Accurate detection and identification of insects from camera trap images with deep learning. *PLOS Sustainability and Transformation*, 2(3):1–18, 2023. 1

[6] Kim Bjerge, Carsten Eie Frigaard, and Henrik Karstoft. Object detection of small insects in time-lapse camera recordings. *Sensors*, 23(16), 2023. 1

[7] Kim Bjerge, Quentin Geissmann, Jamie Alison, Hjalte M.R. Mann, Toke T. Høye, Mads Dyrmann, and Henrik Karstoft. Hierarchical classification of insects with multitask learning and anomaly detection. *Ecological Informatics*, 77:102278, 2023. 1

[8] Kevin F. A. Darras, Marcel Balle, Wenxiu Xu, Yang Yan, Vincent G. Zakka, Manuel Toledo-Hernández, Dong Sheng, Wei Lin, Boyu Zhang, Zhenzhong Lan, Li Fupeng, and Thomas C. Wanger. Eyes on nature: Embedded vision cameras for terrestrial biodiversity monitoring. *Methods in Ecology and Evolution*, 15(12):2262–2275, 2024. 1

[9] Michael S Engel, Luis M P Ceríaco, Gimo M Daniel, Pablo M Dellapé, Ivan Löbl, Milen Marinov, Roberto E Reis, Mark T Young, Alain Dubois, Ishan Agarwal, Pablo Lehmann A., Mabel Alvarado, Nadir Alvarez, Franco Andreone, Katyuscia Araujo-Vieira, John S Ascher, Délio Baêta, Diego Baldo, Suzana A Bandeira, Phillip Barden, Diego A Barrasso, Leila Bendifallah, Flávio A Bockmann, Wolfgang Böhme, Art Borkent, Carlos R F Brandão, Stephen D Busack, Seth M Bybee, Alan Channing, Stylianos Chatzimanolis, Maarten J M Christenhusz, Jorge V Crisci, Guillermo D'elía, Luis M Da Costa, Steven R Davis, Carlos Alberto S De Lucena, Thierry Deuve, Sara Fernandes Elizalde, Julián Faivovich, Harith Farooq, Adam W Ferguson, Spartaco Gippoliti, Francisco M P Gonçalves, Victor H Gonzalez, Eli Greenbaum, Ismael A Hinojosa-Díaz, Ivan Ineich, Jianping Jiang, Sih Kahono, Adriano B Kury, Paulo H F Lucinda, John D Lynch, Valéry Malécot, Mariana P Marques, John W M Marris, Ryan C Mckellar, Luis F Mendes, Silvio S Nihei, Kanto Nishikawa, Annemarie Ohler, Victor G D Orrico, Hidetoshi Ota, Jorge Paiva, Diogo Parrinha, Olivier S G Pauwels, Martín O Pereyra, Lueji B Pestana, Paulo D P Pinheiro, Lorenzo Prendini, Jakub Prokop, Claus Rasmussen, Mark-Oliver Rödel, Miguel Trefaut Rodrigues, Sara M Rodríguez, Hearty Salatnaya, Íris Sampaio, Alba Sánchez-García, Mohamed A Shebl, Bruna S Santos, Mónica M Solórzano-Kraemer, Ana C A Sousa, Pavel Stoev, Pablo Teta, Jean-François Trape, Carmen Van-Dúnem Dos Santos, Karthikeyan Vasudevan, Cor J Vink, Gernot Vogel, Philipp Wagner, Torsten Wappler, Jessica L Ware, Sonja Wedmann, and Chifundera Kusamba Zacharie. The taxonomic impediment: a shortage of taxonomists, not the lack of technical approaches. *Zoological Journal of the Linnean Society*, 193(2):381–387, 2021. 1

[10] Ross J. Gardiner, Sareh Rowlands, and Benno I. Simmons. Towards scalable insect monitoring: Ultra-lightweight cnns as on-device triggers for insect camera traps. *Methods in Ecology and Evolution*. 1

[11] Zeming Gong, Austin T. Wang, Xiaoliang Huo, Joakim B. Haurum, Scott C. Lowe, Graham W. Taylor, and Angel X. Chang. Clibd: Bridging vision and genomics for biodiversity monitoring at scale. *arXiv preprint*, 2025. 2

[12] Jianyang Gu, Samuel Stevens, Elizabeth G. Campolongo, Matthew J. Thompson, Net Zhang, Jiaman Wu, et al. Bioclip 2: Emergent properties from scaling hierarchical contrastive learning. *arXiv preprint*, 2025. 2

[13] Toke T. Høye, Johanna Ärje, Kim Bjerge, Oskar L. P. Hansen, Alexandros Iosifidis, Florian Leese, Hjalte M. R. Mann, Kristian Meissner, Claus Melvad, and Jenni Raitoharju. Deep learning and computer vision will transform entomology. *Proceedings of the National Academy of Sciences*, 118(2):e2002545117, 2021. 1

[14] Aditya Jain, Fagner Cunha, Michael James Bunsen, Juan Sebastián Cañas, Léonard Pasi, Nathan Pinoy, Flemming Helsing, JoAnne Russo, Marc Botham, Michael Sabourin, et al. Insect identification in the wild: The ami dataset. In *European Conference on Computer Vision*, pages 55–73. Springer, 2024. 1, 2, 3

[15] Jiangtao Li, Yuwei Su, Zhaojun Cui, Jida Tian, and Huiling Zhou. A method to establish a synthetic image dataset of stored-product insects for insect detection. *IEEE Access*, 10: 70269–70278, 2022. 1

[16] Yuang Liu, Wei Zhang, Jun Wang, and Jianyong Wang. Data-free knowledge transfer: A survey. *CoRR*, abs/2112.15278, 2021. 2

[17] Zhuang Liu, Hanzi Mao, Chao-Yuan Wu, Christoph Feichtenhofer, Trevor Darrell, and Saining Xie. A convnet for the 2020s. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 11976–11986, 2022. 2

[18] John E. Losey and Mace Vaughan. The economic value of ecological services provided by insects. *BioScience*, 56(4): 311–323, 2006. 1

[19] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017. 3

[20] Graham A. Montgomery, Michael W. Belitz, Rob P. Guralnick, and Morgan W. Tingley. Standards and best practices for monitoring and benchmarking insects. *Frontiers in Ecology and Evolution*, 8, 2021. 1, 2

[21] Amir Moslemi, Anna Briskina, Zubeka Dang, and Jason Li. A survey on knowledge distillation: Recent advancements. *Machine Learning with Applications*, 18:100605, 2024. 2, 3

[22] Hoang-Quan Nguyen, Thanh-Dat Truong, Xuan Bac Nguyen, Ashley Dowling, Xin Li, and Khoa Luu. Insect-foundation: A foundation model and large-scale 1m dataset for visual insect understanding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024. 2

[23] Rita Pucci, Vincent J. Kalkman, and Dan Stowell. Performance of computer vision algorithms for fine-grained classification using crowdsourced insect images. *IET Computer Vision*, 19(1):e70006, 2025. 1

[24] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PmLR, 2021. 2

[25] Miklas Riechmann, Ross Gardiner, Kai Waddington, Ryan Rueger, Frederic Fol Leymarie, and Stefan Rueger. Motion vectors and deep neural networks for video camera traps. *Ecological Informatics*, 69:101657, 2022. 1

[26] D. B. Roy, J. Alison, T. A. August, M. Bélisle, K. Bjerge, J. J. Bowden, M. J. Bunsen, F. Cunha, Q. Geissmann, K. Goldmann, A. Gomez-Segura, A. Jain, C. Huijbers, M. Larrivée, J. L. Lawson, H. M. Mann, M. J. Mazerolle, K. P. McFarland, L. Pasi, S. Peters, N. Pinoy, D. Rolnick, G. L. Skinner, O. T. Strickson, A. Svenning, S. Teagle, and T. T. Høye. Towards a standardized framework for ai-assisted, image-based monitoring of nocturnal insects. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 379(1904):20230108, 2024. 1

[27] Maximilian Sittinger, Johannes Uhler, Maximilian Pink, and Annette Herz. Insect detect: An open-source diy camera trap for automated insect monitoring. *PLOS ONE*, 19(4):1–28, 2024. 1

[28] Samuel Stevens, Jiaman Wu, Matthew J. Thompson, Elizabeth G. Campolongo, Chan Hee Song, David E. Carlyn, et al. Bioclip: A vision foundation model for the tree of life. *arXiv preprint*, 2023. 2

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

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

[31] Roel van Klink, Diana E. Bowler, Konstantin B. Gongalsky, Ann B. Swengel, Alessandro Gentile, and Jonathan M. Chase. Meta-analysis reveals declines in terrestrial but increases in freshwater insect abundances. *Science*, 368(6489): 417–420, 2020. 1

[32] Jiří Vyskočil and Lukas Picek. Towards zero-shot camera trap image categorization. In *European Conference on Computer Vision*, pages 37–53. Springer, 2025. 2

[33] David L. Wagner, Eliza M. Grames, Matthew L. Forister, May R. Berenbaum, and David Stopak. Insect decline in the anthropocene: Death by a thousand cuts. *Proceedings of the National Academy of Sciences*, 118(2):e2023989118, 2021. 1

[34] Mei Wang and Weihong Deng. Deep visual domain adaptation: A survey. *Neurocomputing*, 312:135–153, 2018. 1

[35] Chih-Hsuan Yang, Benjamin Feuer, Zaki Jubery, Zi K. Deng, Andre Nakkab, Md Zahid Hasan, et al. Biotrove: A large curated image dataset enabling ai for biodiversity. *arXiv preprint*, 2024. 2

Bridging Domain Gaps for Fine-Grained Moth Classification Through Expert-Informed Adaptation and Foundation Model Priors

Supplementary Material

8. Embedding-Based Clustering for Train/Test Splitting

To obtain train/test splits from AMI data that are decoupled, we first embed each AMI image using a ResNet-50 network pretrained on ImageNet-1K; this places visually similar images near one another in the embedding space. For every moth class, we then perform agglomerative hierarchical clustering with cosine distance, setting the number of clusters to $\max(K_{\text{min}}, \sqrt{N})$, where N is the number of images in the class and $K_{\text{min}} = 5$. After clustering, we randomly permute the cluster order and walk through this shuffled list, adding images to the test split until the specified amount of test examples are collected. This selective part is illustrated in Figure 4.

Figure 4. List of image embeddings with each showing its cluster number. Cluster groups have been shuffled. Each element are coloured and labelled by a unique cluster ID. The test set is partitioned as the first images in this list, with the remainder as train.

This approach enforces semantic separation between the train/test split, reducing bias and providing a realistic assessment of model generalisation.

9. Dataset Composition

Target-domain Mix (%)	Target-domain (AMI) Contribution	Source-domain (GBIF) Contribution	Total Dataset Size
0%	0	18573	18573
1%	187	18573	18760
5%	997	18573	19550
10%	2063	18573	20636
20%	4643	18573	23216
25%	6191	18573	24764
33%	9147	18573	27720
50%	10100	10100	20200

Table 2. Statistics of each domain-mixed training dataset at various amounts of target-domain percentages. Contributions from the target-domain (AMI) images and the source domain (GBIF) images vary to meet the desired domain mix fraction.