Supplementary
Insect-Foundation: A Foundation Model and Large-scale 1M Dataset for Visual Insect Understanding

Table 1. **Effectiveness of Patch Sampling Ratio.** We evaluate the impact of the sampling ratio on IP102 Classification [5] with four sampling ratio, i.e. 25%, 50%, 75%, and 90%.

Sampling Ratio	25%	**50%**	75%	90%
Accuracy@1 (%)	65.2	**73.3**	72.1	69.8
Accuracy@5 (%)	82.8	**91.6**	90.9	88.3

A. Insect-1M Dataset

The dataset is provided in a JSON file naming `Insect-1M.json`. The JSON file includes two attributes: `insect_records` and `description_records`. Each record in `insect_records` includes insect ID, taxonomies, image URL, and a list of description IDs. Each record in `description_records` includes description ID, taxonomic name, and description.

B. Fine-tuning Details

For the IP102 classification task, the images are rescaled into 256 for the shortest side and randomly cropped into 224×224. The model is fine-tuned iteratively with 30 epochs. The pre-trained backbone of the proposed model, i.e., ViT, is used for fine-tuning. For the IP102 detection task, we use Faster-RCNN [3] with the pre-trained backbone of the proposed model. The images are rescaled randomly from 400 to 800 for the shortest side in the training phase and rescaled into 800 in the testing phase.

C. Effectiveness of Patch Sampling Ratio

Table 1 shows the evaluation results of the patch sampling ratio selection affecting the model performance. It shows that the sampling ratio of 50% is the best ratio when the lower ratio of 25% prevents the model from having sufficient information for pretraining. Meanwhile, higher ratios, i.e., 75% and 90%, weaken the learning ability of the model.

Table 2. **Classification results on iNat-2021 Insect Benchmark [4].** Both proposed models pre-trained with and without the insect descriptions outperform prior methods by a large margin.

Method	Description	Pre-train Data	Acc@1 (%)	Acc@5 (%)
Vit-B/16 [1]	✗	ImageNet1K	87.00	96.21
MAE [2]	✗	Insect-1M	87.52	96.42
CoCa [6]	✓	Insect-1M	88.22	96.70
Insect-Foundation	✗	**Insect-1M**	**89.23**	**96.88**
Insect-Foundation	✓	**Insect-1M**	**90.40**	**97.36**

D. Additional Experimental Results

We have extended experimental results of the Insect subset of iNat-2021 [4] for the classification task on the full training dataset. As shown in Table 2, compared to MAE and CoCa, our method achieves higher accuracy from 87.52% to 89.23% without taxonomic descriptions and from 88.22% to 90.40% with descriptions.

E. Image Sources and Copyright

The Insect-1M dataset is collected from several sources. A major part of this dataset comes from our own photos captured from real insect samples provided by the Entomology and Plant Pathology Department Library of the University of Arkansas from a 5-year research project collaboration related to this work. We own the copyright for this portion of the dataset. Another part of the dataset was collected from various sources on the Internet. We respect the copyright of these images and provide their URLs in the dataset.

References

[1] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020. 1

[2] Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked autoencoders are scalable vision learners. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 16000–16009, 2022. 1

[3] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. *Advances in neural information processing systems*, 28, 2015. 1

[4] Grant Van Horn, Elijah Cole, Sara Beery, Kimberly Wilber, Serge Belongie, and Oisin Mac Aodha. Benchmarking representation learning for natural world image collections. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 12884–12893, 2021. 1

[5] Xiaoping Wu, Chi Zhan, Yu-Kun Lai, Ming-Ming Cheng, and Jufeng Yang. Ip102: A large-scale benchmark dataset for insect pest recognition. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 8787–8796, 2019. 1

[6] Jiahui Yu, Zirui Wang, Vijay Vasudevan, Legg Yeung, Mojtaba Seyedhosseini, and Yonghui Wu. Coca: Contrastive captioners are image-text foundation models. *arXiv preprint arXiv:2205.01917*, 2022. 1