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

Hoang-Quan Nguyen[1]*, Thanh-Dat Truong[1]*, Xuan Bac Nguyen[1], Ashley Dowling[2], Xin Li[3], Khoa Luu[1]
[1]Department of Electrical Engineering and Computer Science, University of Arkansas, AR
[2]Department of Entomology and Plant Pathology, University of Arkansas, AR
[3]Department of Computer Science, SUNY Albany, NY
{hn016, tt032, xnguyen, adowling, khoaluu}@uark.edu, xli48@albany.edu
https://uark-cviu.github.io/projects/insect_foundation.html

Abstract

In precision agriculture, the detection and recognition of insects play an essential role in the ability of crops to grow healthy and produce a high-quality yield. The current machine vision model requires a large volume of data to achieve high performance. However, there are approximately 5.5 million different insect species in the world. None of the existing insect datasets can cover even a fraction of them due to varying geographic locations and acquisition costs. In this paper, we introduce a novel "Insect-1M" dataset, a game-changing resource poised to revolutionize insect-related foundation model training. Covering a vast spectrum of insect species, our dataset, including 1 million images with dense identification labels of taxonomy hierarchy and insect descriptions, offers a panoramic view of entomology, enabling foundation models to comprehend visual and semantic information about insects like never before. Then, to efficiently establish an Insect Foundation Model, we develop a micro-feature self-supervised learning method with a Patch-wise Relevant Attention mechanism capable of discerning the subtle differences among insect images. In addition, we introduce Description Consistency loss to improve micro-feature modeling via insect descriptions. Through our experiments, we illustrate the effectiveness of our proposed approach in insect modeling and achieve State-of-the-Art performance on standard benchmarks of insect-related tasks. Our Insect Foundation Model and Dataset promise to empower the next generation of insect-related vision models, bringing them closer to the ultimate goal of precision agriculture.

1. Introduction

Insects are the most diverse and abundant eukaryotic organisms on the planet. They inhabit all terrestrial and aquatic habitats and play a significant role within their community, habitat, and ecosystem as contributors to nutrient cycling, maintenance of plant and animal communities, disease cycling, and overall ecosystem health. Therefore, in the agricultural revolution, the detection and identification of insects plays a key role in ensuring healthy crop growth and high-quality production. Prior methods [2, 3, 6, 32, 55, 66] often fine-tuned the pre-trained ImageNet models on insect data for specific insect-related tasks, e.g., Insect Classification [2, 6, 13, 66], Insect Detection [66]. However, these methods remained limited since the models pre-trained on ImageNet [12, 15, 16, 20, 50, 52] could not model the micro features of insects, e.g., tiny texture and details of insects, as ImageNet [12] is the generic object dataset.

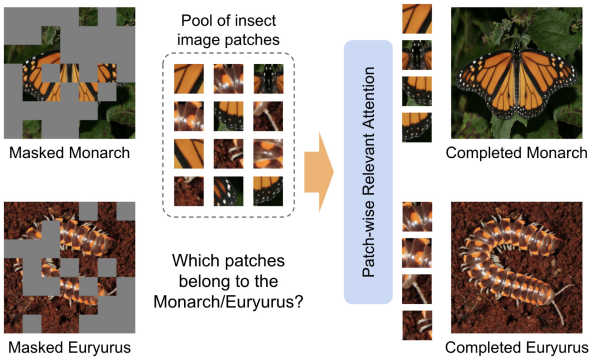

Figure 1. **Our Proposed Patch-wise Relevant Attention.** Given masked insect images and separated image patches, our model can discriminate these patches that have small differences via relevant scores computed between masked images and image patches.

Recent foundation models [7–9, 17–19, 40, 43, 69, 70] pre-trained on large-scale datasets have revolutionized vision models with solid performance on downstream applications. These models are designed to model general or specific properties of images or videos that can later be generalized to downstream tasks and unseen data. The capability of the foundation model is often implemented with

*Co-first authors

Table 1. Comparison with existing datasets related to insects. Our proposed dataset has hierarchical labels with 6 main hierarchical levels, i.e., Subphylum, Class, Order, Family, Genus, and Species, and large numbers of species and samples. Moreover, the proposed dataset contains hierarchical descriptions for each insect and auxiliary taxonomic level, i.e., Subclass, Suborder, Subfamily, etc.

Dataset	Year	Species	Hierarchical Labels	Hierarchical Levels	Insect Description	Auxiliary Taxonomic Level	Number of Samples
Samanta et al. [48]	2012	8	✗	1	✗	✗	609
Wang et al. [61]	2012	221	✓	3	✗	✗	225
Venugoban et al. [60]	2014	20	✗	1	✗	✗	200
Xie et al. [67]	2015	24	✗	1	✗	✗	1,440
Liu et al. [28]	2016	12	✗	1	✗	✗	5,136
Xie et al. [68]	2018	40	✗	1	✗	✗	4,500
Deng et al. [13]	2018	10	✗	1	✗	✗	563
Alfarisy et al. [1]	2018	13	✗	1	✗	✗	4,511
PestNet [25]	2019	16	✗	1	✗	✗	88,670
IP102 [66]	2019	102	✓	3	✗	✗	75,222
AgriPest [63]	2021	14	✓	2	✗	✗	49,707
INSECT [4]	2021	1,213	✗	1	✗	✗	21,212
iNat-2021 [59]	2021	2,752	✓	5	✗	✗	723,816
Our Insect-1M	**2023**	**34,212**	✓	**6**	✓	✓	**1,017,036**

self-supervised or prompt-engineering training on large-scale datasets [12, 19, 49, 71]. However, the current insect datasets [1, 2, 6, 13, 28, 48, 60, 61, 66–68] are insufficient to establish the foundation model of insects due to their scale and diversity. Indeed, the most recent work presents an insect recognition dataset containing over $75,000$ images of 102 species [66]. Although the dataset includes many species, compared to the species of insects in the natural environment with over 5.5 million species [45, 51], the current work needs to have the diversity of insects. Furthermore, to our knowledge, the current insect dataset [66] does not provide the corresponding insect descriptions, limiting the ability to learn the foundation models.

Although the dataset is an important factor in developing an insect foundation model, the learning approach of the foundation model plays a significant role in performance. There is significant progress in developing vision foundation models. Common approaches learned alignment between vision and language, for example, CLIP [43], ALIGN [19], CoCa [70], to model visual concepts and data distributions. Meanwhile, self-supervised contrastive or distillation learning approaches, e.g., MoCo [9, 10, 17], DINO [7, 40], MAE [18], etc., learned the vision model by various pre-text tasks and have shown its scaling ability and generalizes well to various downstream tasks. However, most of these previous foundation models represent the general information of natural images without specific knowledge. When deploying in the insect domains, they cannot capture the micro-features of insects, i.e., key features or appearance to distinguish the species, since the texture and details of insects are often small and diverse compared to generic objects. Meanwhile, fine-grained discrimination between insect images is crucial in insect foundation models due to the high diversity of species. Therefore, to successfully develop the insect foundation model, the learning approach needs to understand and be able to model the micro-features of insects. Based on this observation, we present a novel pre-text task to enhance the recognition ability of the model between small features of the insect, as illustrated in Fig. 1.

Contributions of this Work: To contribute to the development of the Insect Foundation Model in precision agriculture, we introduce a novel large-scale insect dataset, i.e., ***Insect-1M*** , and a new Insect Foundation Model, i.e., ***Insect-Foundation***, that can transfer to various downstream insect-related applications, e.g., insect detection, insect classification, insect vision-language understanding. Our contributions can be summarized as follows. First, we present a new rich and large-volume insect dataset, i.e., ***Insect-1M***, that consists of 1 million images of insects with dense identifications of taxonomy hierarchy from the abstract level of taxonomy, e.g., Class, Order, to the detailed level of taxonomy, e.g., Genus, Species. In addition, each insect contains a detailed description that describes the details and features of insects. To the best of our knowledge, our proposed Insect-1M dataset is $13\times$ larger than the prior published IP102 dataset [66]. Second, to model the micro features of insects, we introduce a new self-supervised contrastive learning paradigm with a novel Patch-wise Relevant Attention mechanism to model the feature correlations of insect details. Third, to increase the modeling capability of the Insect Foundation Model in learning insect details, we introduce a new Description Consistency loss to learn the detailed features of insects via the textual description. Finally, through our intensive experiments on the Insect Classification and Insect Detection benchmarks [66], we show the effectiveness of our approach in insect modeling and our superior performance compared to the prior methods.

2. Related Work

Insect Datasets. There are prior studies releasing insect datasets on a small scale for recognition problems. [60] presented a dataset consisting of 20 species with 10 samples

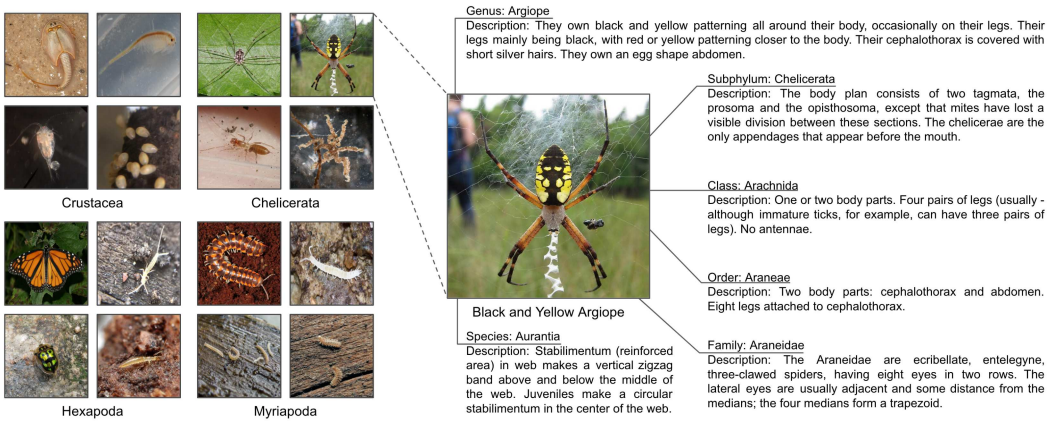

Figure 2. **Examples of Our Insect-1M Dataset.** The left figure illustrates the samples of the four Subphylums, including Chelicerata, Crustacea, Hexapoda, and Myriapoda. The right figure shows an example of hierarchical descriptions of the Aurantia Species.

for each species. Then, [67] introduced an insect dataset including $1,440$ samples of 24 species. Several subsequent studies have larger datasets for deep learning, e.g., [68] proposed an insect dataset of $4,500$ images with 40 different species for insect classification, and [28] proposed an insect dataset with over $5,000$ samples for insect recognition and localization. PestNet [25] and AgriPest [63] were introduced for the small pest detection task. Recently, [66] has presented IP102 as a large-scale dataset containing over $75K$ samples of insects with 102 species for classification and detection tasks. Meanwhile, [59] proposed a large-scale dataset including over $723K$ samples of Arthropoda phylum with $2,752$ species. Although prior efforts promoted the development of vision and machine intelligence in precision agriculture, no dataset has a large volume of samples and diverse species for insect-related foundation model training. Therefore, this work introduces a novel dataset that not only contains a large number of samples, i.e. 1M images, but also has hierarchical labels from the high to the low taxonomy level, including class, order, family, genus, and species. Table 1 compares our proposed dataset with the prior ones. In comparison with prior datasets, the number of images in our proposed Insect-1M dataset is $13\times$ higher than the prior IP102 dataset, and the number of species is $335\times$ higher than IP102 [66]. To preserve the rights of datasets and authors of images, instead of publishing images, we only provide labels and links to download images.

Self-supervised Pre-training. Self-supervised pre-training has become a popular strategy for solving visual recognition problems, including classification, localization, segmentation, video recognition, tracking, and many other problems [18, 33–38, 53, 54, 56–58]. SimCLR [8] learned the visual representation of images via a contrastive learning framework using different data augmentation operations. MoCo [17] introduced momentum updating for the encoder while learning the image representation via contrastive learning. The MoCo framework was later used to improve the SimCLR approach without requiring a large training batch size [9]. MoCo-V3 [10] improved prior Momentum Contrastive frameworks by eliminating the memory queue to stabilize the training when the batch size is large. DINO [7] proposed a self-supervised learning approach using knowledge distillation with no labels. Later, it was extended to DINO-V2 [40] by stabilizing self-supervised learning when scaling the size of models and data. BEiT [5] proposed a masked image modeling task and used discrete visual tokens from the original image as prediction targets. MAE [18] and SimMIM [69] directly used a decoder to reconstruct pixel values from masked regions. Jigsaw-ViT [11] presented a pre-training task for transformer models by solving the shuffled patches of images. This learning strategy was also applied on the temporal dimension to improve the robustness of video modeling [54]. Micron-BERT [36] studied the micro-changing in facial videos by learning to detect the minor differences in an image that has swapping regions between two frames.

Joint Vision-Language Pre-training. Recent work introduced joint vision-language pre-training. CLIP [43], and ALIGN [19] addressed that dual-encoder models pre-trained on image-text pairs in contrastive objectives can learn strong representations of image and text for cross-modal alignment and zero-shot image recognition problems. LiT [72] and BASIC [42] proposed zero-shot transfer learning approaches by teaching the text model to learn the representation of the pre-trained image model via contrastive losses with large-scale data. SimVLM [65], OFA [62], and BLIP [22] trained an encoder-decoder model with language generative losses and achieved high performance in the vision-language benchmarks. CoCa [70] utilized contrastive learning and generative image captioning for global representation learning and fine-grained image-text alignment. Later work [73] used sigmoid loss to compute the image-text similarity for batch size scaling. LexLIP [31] projected images into a lexicon space for image-text sparse matching. Meanwhile, EQSIM [64] computed the similarity by the image-text equivariant changing.

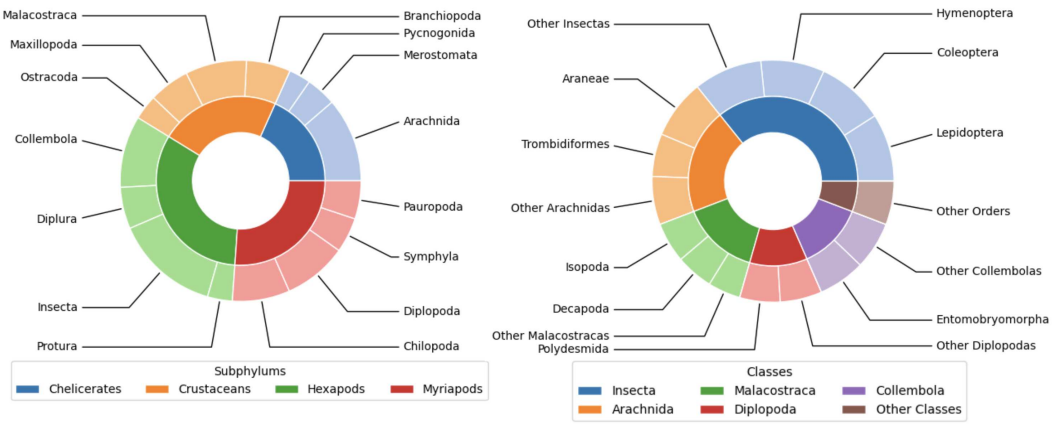

Figure 3. The Distribution of Subphylum and Its Classes (Left) and The Distribution of Class and Its Orders (Right). **Best viewed in color.**

3. The Proposed Insect 1M Dataset

To contribute to establishing the insect foundation model, the large-scale dataset of insects with diverse species is essential. Therefore, we collect a new insect dataset with dense labels of a hierarchical taxonomy. In particular, our Insect-1M dataset contains 1 million insect images with dense hierarchical labels with six main taxonomies, i.e., Subphylum, Class*, Order, Family, Genus, and Species. The samples are in the Phylum Arthropoda and can be divided into 4 Subphylums, which are Chelicerata, Crustacea, Hexapoda, and Myriapoda as shown in Fig. 2. Compared to prior datasets, our Insect-1M has more hierarchical levels with large numbers of species and samples as in Table 1.

3.1. Data Collection Protocol

We utilize insect information containing insect data with images and taxonomies collected by naturalists and entomologists. Each insect sample has a corresponding image and its taxonomic label. From the taxonomic label, we crawl the identification description of the corresponding taxonomy. Notice that the taxonomic labels are hierarchical. The description is written from high-level descriptions, e.g., Subphylum and Class, to low-level descriptions, e.g., Species. Fig. 2 shows an example of an insect description.

3.2. Data Preprocessing and Statistic

Data Preprocessing. The raw data is stored in over 1 million HTML files with predefined HTML structures. Then, we parse the data structures to collect the insect images and their labels. More than 2 million raw images and their corresponding labels have been collected. However, the raw data collected consists of a lot of noise, e.g., incorrect identification of insects, corrupted images, and non-insect images. Therefore, to filter these outliers, our entomology experts must verify the images and their labels, i.e., insect identification. Finally, our collected Insect-1M dataset consists of $1,017,036$ clean images with dense labels of $34,212$ different insect species.

Data Statistic Fig. 3 shows the sample distributions of the Subphylums and their Classes. It is shown that the Class Insecta has the majority of samples. Fig. 3 also illustrates the distribution of the Orders in the major Classes. For each major Class, the data distribution of Orders is well-balanced.

*In this paper, we use the term "Class" as a biological taxonomic level.

4. The Proposed Insect Foundation Model

4.1. Limitations of Prior Foundation Training Approaches

Limitations One of the issues in the visual insect understanding problem is the visual representation and discrimination of the small and undistinguished features of the insects. While MAE [18] reconstructs an image from a masked image for visual representation learning, it focuses

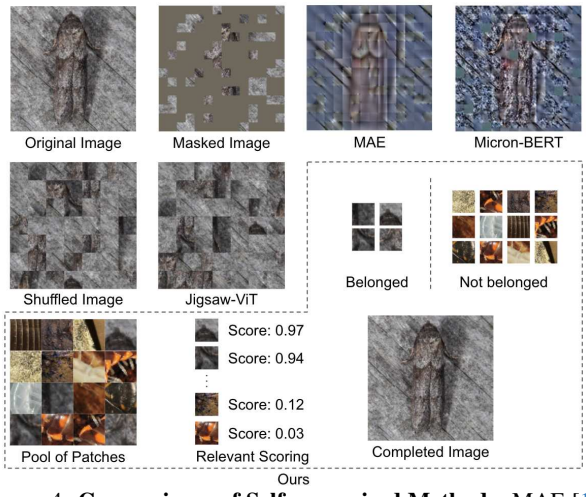

Figure 4. **Comparisons of Self-supervised Methods.** MAE [18] fails to reconstruct the details of the insect since it learns general information about the image. Micron-BERT [36] hardly distinguishes the insect and background. Jigsaw-ViT [11] cannot correct shuffled patches due to confusion between the background and the object. Meanwhile, our approach can find separated patches belonging to the insect by scoring each patch. **Best viewed in color.**

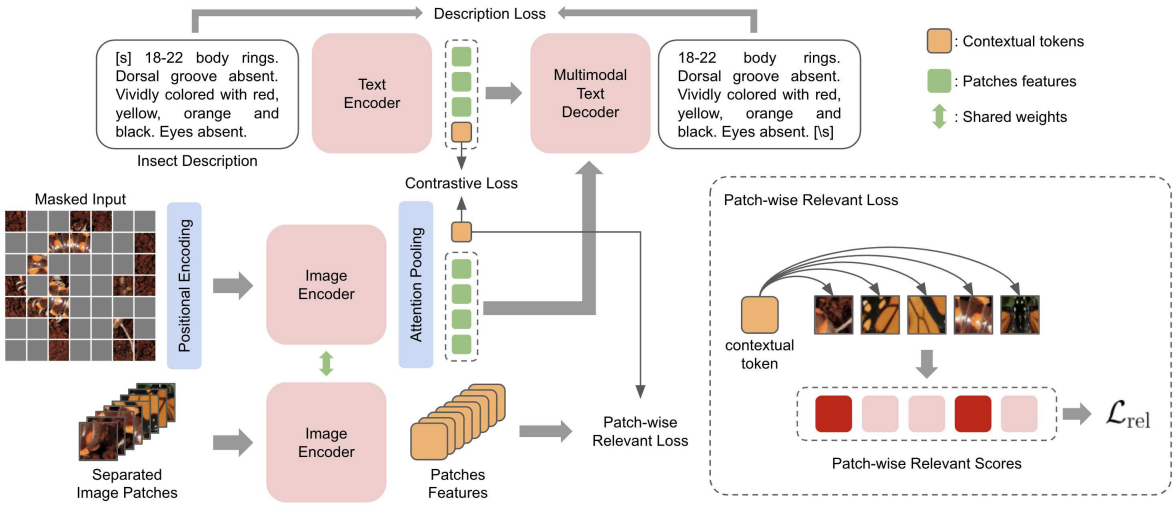

Figure 5. The Overview Framework of Our Proposed Approach to Insect Foundation Model.

on the context inside the image individually without realizing the small details to discriminate between the insects. Meanwhile, Jigsaw solving methods [11, 39] correct the position of image patches to enhance the model robustness to the image structure. This strategy needs more mechanisms to focus on the small details of the image. Micron-BERT [36] highlights the small changes in the image by swapping the regions between two images with similar contexts. However, the small changes in the insect image still preserve the signature features representing the insect. Thus, it makes the model collapse in detecting the small features of insects. Therefore, to address these limitations, we introduce a new approach that learns to recognize the tiny features in the insect images. These features are distinguished from the background by discriminating the minor differences between patches of images individually. Fig. 4 compares prior self-supervised methods [11, 18, 36] with our approach.

Fig. 5 illustrates our insect foundation model. The model is designed to capture the small differences in insect features, i.e., textures or limbs, via our new self-supervised pre-text task. Moreover, the model is pre-trained to learn the fine-grained alignment between the insect description and its visual features. Formally, given an input image I, we divide I into non-overlapping patches. Then, a subset of patches P_s is sampled, and the remaining patches are put into a pool of image patches P_{pool}. The sampling is processed randomly in a uniform distribution. An image encoder is used to map I_p into latent vectors. Given an insect description T of the image, a text encoder is presented to extract information from T. A text decoder and joint image-text contrastive learning module are introduced to map the description into the image. Finally, a Patch-wise Relevant Attention module is proposed for self-supervised learning to enhance the discrimination robustness of the model.

4.2. Input Modeling

An input image $I \in \mathbb{R}^{H\times W\times 3}$ is divided into non-overlapping patches $P = \{p_s^i\}_{i=1}^{N_P}$ where H, W are the height and width of the input image, $N_P = HW/(s_p)^2$ is the number of patches. Each patch p_s^i has a resolution of $s_p \times s_p$. The non-overlapping patches P are then randomly sampled into a subset of patches $P_s \subset P$ and put the other patches into a pool of image patches P_{pool}. Note that P_{pool} contains patches from multiple images in the training set.

4.3. Image Encoder

Each patch $p_s^i \in P_s$ is projected into a latent vector $\mathbf{x}_s^i \in \mathbb{R}^d$ where d is the dimension of the latent vectors. A subset patches P_s can be represented as follows:

$$\mathbf{X}_s = \text{concat}[\mathbf{x}_s^i]_{i=1}^{N_{P_s}} \in \mathbb{R}^{N_{P_s}\times d}, \quad \mathbf{x}_s^i = \alpha_p(p_s^i) + \mathbf{e}_p(i) \quad (1)$$

where α_p and $\mathbf{e}_p$ are the projection embedding and position embedding.

Let an image encoder $E_{\text{image}}(\mathbf{X}_s)$ be a stack of L_e transformer blocks where each block contains multi-head self-attention (MSA) and multi-layer perceptron (MLP).

$$\begin{aligned} \mathbf{X}'_l &= \mathbf{X}_{l-1} + \text{MSA}(\text{LN}(\mathbf{X}_{l-1})) \\ \mathbf{X}_l &= \mathbf{X}'_l + \text{MLP}(\text{LN}(\mathbf{X}'_l)) \\ \mathbf{X}_0 &= \mathbf{X}_s,\ 1 \le l \le L_e \end{aligned} \quad (2)$$

where LN is the layer normalization. Then, given $\mathbf{X}_s$, the output latent vector $\mathbf{Z}_s$ is represented as follows:

$$\mathbf{Z}_s = E_{\text{image}}(\mathbf{X}_s), \quad \mathbf{Z}_s \in \mathbb{R}^{N_{P_s}\times d} \quad (3)$$

4.4. Insect Micro-feature Self-supervised Learning

The recognition of insects relies on the insect texture, eyes, or limbs that are tiny to detect. To make the model robust to the small features of insect images, we propose a self-supervised learning strategy to spot these small features via the small differences in the images. Notice that the insects can be distinguished by detecting and discriminating the

critical features in each part of those insects. To enhance this ability for the model, a pre-text task is presented. In particular, after extracting global information from a masked image of the insect, the vision model learns to find the remaining patches of the image by comparing image patches of different insect species. Thanks to our learning mechanism, the model learns the key features representing each insect and discriminates the small features between different species. As illustrated in Fig. 6, given a subset of patches P_s from the image I and a pool of image patches P_{pool}, we train the model to find the patches $p_t \in P_{\text{pool}}$ that originally belong to the image I. Then, given latent vectors $\mathbf{Z}_s$ of P_s, a patch-wise relevant attention score (PRS) is computed between $\mathbf{Z}_s$ and each patch $p \in P_{\text{pool}}$. The score can be defined as:

$$\text{PRS} = f(\mathbf{Z}_s, p) \in [0, 1] \tag{4}$$

The higher the score is, the more possibility that $p \in P$.

Attention Pooling To compute the relevance between latent vectors $\mathbf{Z}_s$ from the image I and the patch $p \in P_{\text{pool}}$, the latent vectors $\mathbf{Z}_s$ should be aggregated to represent the holistic information of I. Inspired by [70], we compute the global information of I via attention pooling. Given a placeholder contextual token $\mathbf{z}'_{ct}$ as a query $\mathbf{Q}_{ct}$ and latent vectors $\mathbf{Z}_s$ as a key $\mathbf{K}_Z$ and a value $\mathbf{V}_Z$, we compute an attention map between $\mathbf{Q}_{ct}$ and $\mathbf{K}_Z$. Then, a contextual token $\mathbf{z}_{ct}$ representing the global information of I is computed via the attention map and the value $\mathbf{V}_Z$. The attention pooling (Fig. 7) can be formulated as Eqn. (5).

$$\begin{gathered}\mathbf{Q}_{ct} = \text{Linear}(\mathbf{z}'_{ct}) \quad \mathbf{K}_Z = \text{Linear}(\mathbf{Z}_s) \quad \mathbf{V}_Z = \text{Linear}(\mathbf{Z}_s) \\ \mathbf{z}_{ct} = \text{softmax}\left(\frac{\mathbf{Q}_{ct}\mathbf{K}_Z^T}{\sqrt{d}}\right)\mathbf{V}_Z\end{gathered} \tag{5}$$

Patch-wise Relevant Attention Given $\mathbf{z}_{ct}$ as a contextual token representing the information of I, we compute the relevance between $\mathbf{z}_{ct}$ and $p \in P_{\text{pool}}$. From Eqn. (4), we expand the attention score function f as in Eqn. (6).

$$\text{PRS} = f(\mathbf{Z}_s, p) = H(\mathbf{z}_{ct}, \mathbf{z}_p) \tag{6}$$

where $\mathbf{z}_p = E_{\text{image}}(\alpha_p(p))$ is a latent vector representing the patch p, H is a similarity function between two latent vectors. From Eqn. (6), we expand the score function into a self-supervised loss function $\mathcal{L}_{\text{PRS}}$ as follow:

$$\mathcal{L}_{\text{rel}} = -y \log(H(\mathbf{z}_{ct}, \mathbf{z}_p)) - (1-y)\log(1 - H(\mathbf{z}_{ct}, \mathbf{z}_p)) \tag{7}$$

where $y = 1$ if $p \in P$ and $y = 0$ otherwise.

4.5. Fine-grained Insect Image-Text Alignment

Each species has an individual definition and description that can be aligned to parts of the insect image. We adopt a text decoder to generate the species descriptions from insect images. Moreover, to capture the general information of species, we utilize contrastive learning between global features of the insect images and description. As a result, the model can learn specific information from insect images via insect descriptions.

Formally, an insect description text is tokenized into $T = \{t_i\}_{i=1}^{N_T}$ where N_T is the number of tokens of the description. Each token $t_i \in T$ is embedded into a latent vector $\mathbf{w}_i \in \mathbb{R}^d$. The description can be represented as:

$$\mathbf{W} = \text{concat}[\mathbf{w}_i]_{i=1}^{N_T} \in \mathbb{R}^{N_T \times d}, \quad \mathbf{w}_i = \alpha_w + \mathbf{e}_w(i) \tag{8}$$

where α_w and $\mathbf{e}_w$ are the projection embedding and position embedding.

Similar to the image encoder, let the text encoder $E_{\text{text}}(\mathbf{W})$ be a stack of L'_e transformer blocks containing multi-head self-attention and multi-layer perceptron. The output latent vector $\mathbf{Z}'$ of the description is computed as

$$\mathbf{W}' = E_{\text{text}}(\mathbf{W}), \quad \mathbf{Z}' \in \mathbb{R}^{N_T \times d} \tag{9}$$

We then use the latent vector $\mathbf{Z}_s$ of the insect image and $\mathbf{W}'$ of the description text for image-text contrastive learning and multi-modal image description decoding.

Image-text Contrastive Learning. Inspired by the prior language model frameworks [14, 21, 27, 44], a contextual token $\mathbf{w}_{ct}$ representing the semantic information of the description is added at the beginning of $\mathbf{W}$ as in Eqn. 8. Then the two encoders E_{image} and E_{text} can be jointly optimized via contrastive learning as follow:

$$\mathcal{L}_{\text{con}} = \frac{-1}{N}\sum_{i=1}^{N}\left[\log\frac{\exp(\mathbf{z}_i^T\mathbf{w}_i)}{\sum_{j=1}^{N}\exp(\mathbf{z}_i^T\mathbf{w}_j)} + \log\frac{\exp(\mathbf{w}_i^T\mathbf{z}_i)}{\sum_{j=1}^{N}\exp(\mathbf{w}_i^T\mathbf{z}_j)}\right] \tag{10}$$

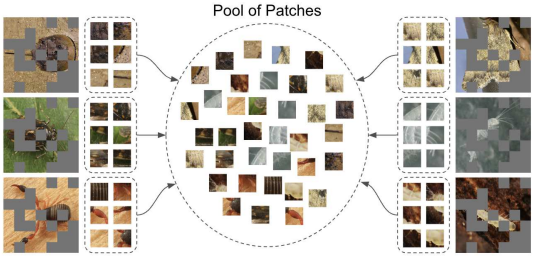

Figure 6. **Pool of Image Patches.** A subset of patches of an image is sampled for image encoding while the remaining patches are placed into a pool of patches for the self-supervised pre-text task.

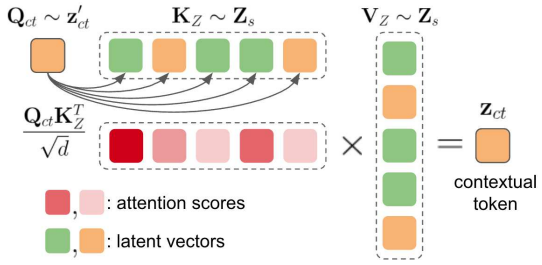

Figure 7. **Attention Pooling Module.** The contextual token $\mathbf{z}_{ct}$ represents the global information of the image I.

Table 2. **Effectiveness of our method on the IP102 Classification.** We evaluate approach with three different vision transformer backbones, i.e., ViT-small/16, ViT-base/16, and ViT-large/16, without or with Attention Pooling (Attn Pool), and three different losses, i.e. Patch-wise Relevant Loss ($\mathcal{L}_{\text{rel}}$), Image-Text Contrastive Loss ($\mathcal{L}_{\text{con}}$), and Description Loss ($\mathcal{L}_{\text{desc}}$).

Backbone	$\mathcal{L}_{\text{rel}}$	Attn Pool	$\mathcal{L}_{\text{con}}$	$\mathcal{L}_{\text{desc}}$	Acc@1 (%)	Acc@5 (%)
ViT-small/16	✓				68.9	88.8
	✓	✓			69.5	89.7
	✓	✓	✓		70.7	**89.9**
	✓	✓	✓	✓	**71.5**	87.7
ViT-base/16	✓				72.4	91.0
	✓	✓			73.3	91.6
	✓	✓	✓		74.2	91.9
	✓	✓	✓	✓	**75.8**	**92.1**
ViT-large/16	✓				73.8	90.9
	✓	✓			74.6	91.6
	✓	✓	✓		75.9	91.4
	✓	✓	✓	✓	**76.9**	**92.7**

where $\mathbf{z}_i$ and $\mathbf{w}_i$ is the contextual token of the i-th insect image and description.

Multi-modal Image Description Decoding. While image-text contrastive learning represents the global semantic information between the image and description, the multi-model image description decoding aims for the fine-grained details by predicting the tokenized texts of T in an autoregressive manner, as shown in Eqn. (11).

$$\mathcal{L}_{\text{desc}} = -\sum_{t=1}^{N_T} \log D_{\text{multi}}(\mathbf{w}_t|\mathbf{W}_{0:t-1}, \mathbf{Z}_s) \tag{11}$$

where D_{multi} is an autoregressive multi-modal text decoder.

5. Experimental Results

5.1. Foundation Model Pre-training

Our experiments use ViT-Base (ViT-B/16) [15] as the backbone. The images are resized and cropped randomly into the resolution of 224×224. Then, each image is divided into patches of 16×16, creating $N_P = 196$ patches. The patch sampling ratio is selected as 50%, and the remaining patches are put into the pool of image patches. Each patch is projected to latent space of $d = 768$ dimensions. The text encoder and multi-modal text decoder are adopted from the pre-trained BERT model [14]. The model is implemented in PyTorch [41] and trained by $16 \times$ A100 GPUs. The learning rate is initially set to 1.5×10^{-4} with the Consine learning rate scheduler [29]. The model is optimized by AdamW [30] with 200 epochs and a batch size of 64 per GPU.

5.2. Datasets and Benchmarks

IP102 Classification [66] provides 102 species of insects and contains 45,095 training samples, 7,508 validation samples, and 22,619 testing samples. For each species, an image might contain a single insect, multiple insects, or even a diseased crop caused by the species. The insects are in different forms for each class, e.g., egg, larva, pupa, and adult. The performance of insect classification is evaluated by the accuracy of Top 1 (Acc@1) and Top 5 (Acc@5).

Table 3. **Classification results on IP102 Classification benchmark.** 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 (%)
ResNet [66]	✗	ImageNet1K	49.4	-
EfficientNet [6]	✗	ImageNet1K	60.7	-
DenseNet [32]	✗	ImageNet1K	61.9	-
GAEnsemble [3]	✗	ImageNet1K	67.1	-
ViT [15]	✗	ImageNet1K	71.6	87.7
MoCo [17]	✗	1M-Insect	70.6	88.4
DINO [7]	✗	1M-Insect	71.5	91.4
MAE [18]	✗	1M-Insect	72.0	91.5
CoCa [70]	✓	1M-Insect	72.8	91.1
Insect-Foundation	✗	1M-Insect	**73.3**	**91.6**
Insect-Foundation	✓	1M-Insect	**75.8**	**92.1**

IP102 Detection [66] includes 15,178 training images and 3,798 testing images of 102 different species. Following the COCO benchmark [23], the insect detection performance is measured by the Average Precision (AP) and Average Precision at IoU thresholds of 0.5 ($\text{AP}^{\cdot 50}$) and 0.75 ($\text{AP}^{\cdot 75}$).

5.3. Ablation Studies

Our ablation experiments study the effectiveness of our proposed model and hyper-parameters on the IP102 Classification Benchmark as shown in Table 2.

Effectiveness of Network Backbones Table 2 studies the impact of different Vision Transformer backbone sizes, including ViT-small/16, ViT-base/16, and ViT-large/16. As shown in our results, the powerful backbone carries more improvement. In particular, when changing the Transformer backbone size from small to base, the accuracy score increases by a large margin of 4.3% while the large Transformer backbone improves the accuracy score by 1.1%.

Effectiveness of Attention Pooling We evaluate the impact of the attention pooling in the visual representation of the insect images. As shown in Table 2, the Attention Pooling has better representation than the standard classification token computed through transformer layers. In particular, the top-1 accuracies for the three backbones, i.e., small, base, and large, have been increased from 68.9% to 69.5%, from 72.4% to 73.3%, and from 73.8% to 74.6%.

Effectiveness of Image-Text Contrastive Loss As reported in Table 2, the model can understand the insect images better when the model learns to match the images and their descriptions. In detail, the accuracy scores have been increased by 0.8%, 0.9%, and 1.3% for the three backbones when applying the Image-Text Contrastive Loss.

Effectiveness of Description Loss The full configuration in Table 2 shows the experimental results of our model using the Description Loss. As shown in Table 2, the Description Loss helps the model to well-align the information between images and the details of descriptions. Hence, the model

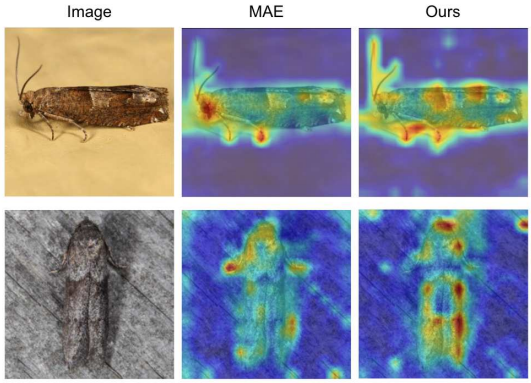

Figure 8. **Attention Visualization.** Compared to MAE [18], our model is robust to small details of insect images. The model can focus on the small textures of the insect, even if the texture is the same as the background (bottom images). **Best viewed in color.**

can represent the fine-grained features of the insects better. In particular, the accuracy scores have been improved from 70.7% to 71.5%, from 74.2% to 75.8%, and from 75.9% to 76.9% for ViT-small/16, ViT-base/16, and ViT-large/16.

5.4. Comparisons with Prior SOTA Methods

Insect Classification Tasks. We fine-tune the linear layer with our pre-trained model on the IP102 dataset [66] for the classification task. As shown in Table 3, our model outperforms deep learning models [15, 16, 20, 50, 52] pre-trained on ImageNet [12] by a large margin. Compared to other pre-training methods [7, 17, 18, 70] on the proposed 1M-Insect dataset, our model shows better performance for both training without and with insect descriptions of 73.3% and 75.8%, respectively. It is shown that the proposed approach has a better visual representation of insect images than the prior pre-training methods on the same dataset.

Visualization Results Fig. 8 visualizes the attention maps of our model compared to MAE [18] pre-trained on the proposed dataset. Since the textures are similar to the background, it is hard for MAE to focus on the small details of the insect. On the contrary, our model can detect the key features, i.e., the textures and the limbs, of the insects.

Zero-shot Insect Classification. We evaluate the performance of our model on the IP102 dataset [66] in a zero-shot manner. In detail, a description corresponds to each species to make the text encoder extract more semantic information about each species. Then, for each insect image, we use the image encoder to extract global features and compare them to each description feature to predict the insect species. Table 4 reports the results of zero-shot classification on the IP102 Classification benchmark. Our model outperforms prior image-text pre-training methods [43, 70, 72] at an accuracy of 49.9%. It shows that our model has well-alignment between the insect image and its description.

Table 4. **Zero-shot classification results on IP102 Classification benchmark.** The proposed model outperforms prior vision-language pretraining methods.

Method	Pretrain Data	Accuracy (%)
CLIP [43]	1M-Insect	41.1
LiT [72]	1M-Insect	43.6
CoCa [70]	1M-Insect	45.3
Insect-Foundation	1M-Insect	**49.9**

Table 5. **Detection results on IP102 Detection benchmark.** The proposed model outperforms prior pre-training methods.

Method	Backbone	Pre-train Data	AP (%)	AP^{50} (%)	AP^{75} (%)
FRCNN [47]	VGG-16 [50]	ImageNet1K	21.1	47.9	15.2
FPN [24]	ResNet-50 [16]	ImageNet1K	28.1	54.9	23.3
SSD300 [26]	VGG-16 [50]	ImageNet1K	21.5	47.2	16.6
RefineDet [74]	VGG-16 [50]	ImageNet1K	22.8	49.0	16.8
YOLOv3 [46]	DarkNet-53 [46]	ImageNet1K	25.7	50.6	21.8
FPN [24]	ViT [15]	ImageNet1K	32.8	54.7	35.0
FPN [24]	MoCo [17]	1M-Insect	33.6	56.1	35.3
FPN [24]	DINO [7]	1M-Insect	34.0	55.8	37.1
FPN [24]	MAE [18]	1M-Insect	34.7	58.4	37.8
FPN [24]	Insect-Foundation	1M-Insect	**36.6**	**59.1**	**40.3**

Insect Detection Tasks. As shown in Table 5, we train a Faster R-CNN model [47] on the IP102 Detection dataset with the ViT backbone adapted for FPN [24]. Compared to models pre-trained on ImageNet [12], our model achieves SOTA results with an average precision of 36.6% and AP^{50} of 59.1% higher than the same backbone pre-trained on ImageNet [12] having AP of 32.8% and AP^{50} of 54.7%. Compared to other self-supervised methods [7, 17, 18], our model achieves higher precision. Thus, our model focuses on the features of insects better than prior methods.

6. Conclusions

This paper has introduced a new large-scale Insect-1M dataset that supports the development of the Insect Foundation Model in precision agriculture. Our proposed dataset includes a large diversity of insect species and multi-level labels of taxonomy. In addition, Insect-1M consists of detailed descriptions of insects that support vision-language insect model training. Then, to improve the micro-feature modeling of our insect foundation model, we introduce a new Patch-wise Relevant Attention mechanism and Description Consistency loss to learn the details of insects. Our experimental results have illustrated the effectiveness and significance of our Insect-1M and Insect Foundation Model.

Limitations This study used a specific network design and learning hyper-parameter to support our hypothesis. However, our approach potentially consists of several limitations related to the design of our Patch-wise Relevant Attention mechanism, where the patches of background and foreground are equally treated. It could result in difficulty in learning the different features of insects. This limitation will further motivate future research to improve the Insect Foundation Model and Micro-feature Modeling.

Acknowledgment. This work is partly supported by NSF DART, NSF SBIR Phase 2, and JBHunt Company. We also acknowledge the Arkansas High-Performance Computing Center for GPU servers and Jesse Ford for dataset tasks.

References

[1] Ahmad Arib Alfarisy, Quan Chen, and Minyi Guo. Deep learning based classification for paddy pests & diseases recognition. In *Proceedings of 2018 international conference on mathematics and artificial intelligence*, pages 21–25, 2018. 2

[2] Adão Nunes Alves, Witenberg SR Souza, and Díbio Leandro Borges. Cotton pests classification in field-based images using deep residual networks. *Computers and Electronics in Agriculture*, 174:105488, 2020. 1, 2

[3] Enes Ayan, Hasan Erbay, and Fatih Varçın. Crop pest classification with a genetic algorithm-based weighted ensemble of deep convolutional neural networks. *Computers and Electronics in Agriculture*, 179:105809, 2020. 1, 7

[4] Sarkhan Badirli, Zeynep Akata, George Mohler, Christine Picard, and Mehmet M Dundar. Fine-grained zero-shot learning with dna as side information. *Advances in Neural Information Processing Systems*, 34:19352–19362, 2021. 2

[5] Hangbo Bao, Li Dong, Songhao Piao, and Furu Wei. Beit: Bert pre-training of image transformers. *arXiv preprint arXiv:2106.08254*, 2021. 3

[6] Edson Bollis, Helio Pedrini, and Sandra Avila. Weakly supervised learning guided by activation mapping applied to a novel citrus pest benchmark. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pages 70–71, 2020. 1, 2, 7

[7] Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 9650–9660, 2021. 1, 2, 3, 7, 8

[8] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *International conference on machine learning*, pages 1597–1607. PMLR, 2020. 3

[9] Xinlei Chen, Haoqi Fan, Ross Girshick, and Kaiming He. Improved baselines with momentum contrastive learning. *arXiv preprint arXiv:2003.04297*, 2020. 1, 2, 3

[10] X Chen, S Xie, and K He. An empirical study of training self-supervised vision transformers. in 2021 ieee. In *CVF International Conference on Computer Vision (ICCV)*, pages 9620–9629, 2021. 2, 3

[11] Yingyi Chen, Xi Shen, Yahui Liu, Qinghua Tao, and Johan AK Suykens. Jigsaw-vit: Learning jigsaw puzzles in vision transformer. *Pattern Recognition Letters*, 166:53–60, 2023. 3, 4, 5

[12] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009. 1, 2, 8

[13] Limiao Deng, Yanjiang Wang, Zhongzhi Han, and Renshi Yu. Research on insect pest image detection and recognition based on bio-inspired methods. *Biosystems Engineering*, 169:139–148, 2018. 1, 2

[14] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018. 6, 7

[15] 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, 7, 8

[16] 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*, pages 770–778, 2016. 1, 8

[17] 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*, pages 9729–9738, 2020. 1, 2, 3, 7, 8

[18] 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. 2, 3, 4, 5, 7, 8

[19] Chao Jia, Yinfei Yang, Ye Xia, Yi-Ting Chen, Zarana Parekh, Hieu Pham, Quoc Le, Yun-Hsuan Sung, Zhen Li, and Tom Duerig. Scaling up visual and vision-language representation learning with noisy text supervision. In *International conference on machine learning*, pages 4904–4916. PMLR, 2021. 1, 2, 3

[20] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. *Advances in neural information processing systems*, 25, 2012. 1, 8

[21] Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Ves Stoyanov, and Luke Zettlemoyer. Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. *arXiv preprint arXiv:1910.13461*, 2019. 6

[22] Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *International Conference on Machine Learning*, pages 12888–12900. PMLR, 2022. 3

[23] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V 13*, pages 740–755. Springer, 2014. 7

[24] Tsung-Yi Lin, Piotr Dollár, Ross Girshick, Kaiming He, Bharath Hariharan, and Serge Belongie. Feature pyramid networks for object detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2117–2125, 2017. 8

[25] Liu Liu, Rujing Wang, Chengjun Xie, Po Yang, Fangyuan Wang, Sud Sudirman, and Wancai Liu. Pestnet: An end-to-end deep learning approach for large-scale multi-class pest detection and classification. *Ieee Access*, 7:45301–45312, 2019. 2, 3

[26] Wei Liu, Dragomir Anguelov, Dumitru Erhan, Christian Szegedy, Scott Reed, Cheng-Yang Fu, and Alexander C Berg. Ssd: Single shot multibox detector. In *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11–14, 2016, Proceedings, Part I 14*, pages 21–37. Springer, 2016. 8

[27] Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*, 2019. 6

[28] Ziyi Liu, Junfeng Gao, Guoguo Yang, Huan Zhang, and Yong He. Localization and classification of paddy field pests using a saliency map and deep convolutional neural network. *Scientific reports*, 6(1):20410, 2016. 2, 3

[29] Ilya Loshchilov and Frank Hutter. Sgdr: Stochastic gradient descent with warm restarts. *arXiv preprint arXiv:1608.03983*, 2016. 7

[30] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017. 7

[31] Ziyang Luo, Pu Zhao, Can Xu, Xiubo Geng, Tao Shen, Chongyang Tao, Jing Ma, Qingwei Lin, and Daxin Jiang. Lexlip: Lexicon-bottlenecked language-image pre-training for large-scale image-text sparse retrieval. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 11206–11217, 2023. 3

[32] Loris Nanni, Gianluca Maguolo, and Fabio Pancino. Insect pest image detection and recognition based on bio-inspired methods. *Ecological Informatics*, 57:101089, 2020. 1, 7

[33] Xuan-Bac Nguyen, Guee Sang Lee, Soo Hyung Kim, and Hyung Jeong Yang. Self-supervised learning based on spatial awareness for medical image analysis. *IEEE Access*, 8: 162973–162981, 2020. 3

[34] Xuan-Bac Nguyen, Duc Toan Bui, Chi Nhan Duong, Tien D Bui, and Khoa Luu. Clusformer: A transformer based clustering approach to unsupervised large-scale face and visual landmark recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10847–10856, 2021.

[35] Xuan Bac Nguyen, Apoorva Bisht, Hugh Churchill, and Khoa Luu. Two-dimensional quantum material identification via self-attention and soft-labeling in deep learning. *arXiv preprint arXiv:2205.15948*, 2022.

[36] Xuan-Bac Nguyen, Chi Nhan Duong, Xin Li, Susan Gauch, Han-Seok Seo, and Khoa Luu. Micron-bert: Bert-based facial micro-expression recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1482–1492, 2023. 3, 4, 5

[37] Xuan-Bac Nguyen, Chi Nhan Duong, Marios Savvides, Kaushik Roy, and Khoa Luu. Fairness in visual clustering: A novel transformer clustering approach. *arXiv preprint arXiv:2304.07408*, 2023.

[38] Xuan-Bac Nguyen, Xin Li, Samee U Khan, and Khoa Luu. Brainformer: Modeling mri brain functions to machine vision. *arXiv preprint arXiv:2312.00236*, 2023. 3

[39] Mehdi Noroozi and Paolo Favaro. Unsupervised learning of visual representations by solving jigsaw puzzles. In *European conference on computer vision*, pages 69–84. Springer, 2016. 5

[40] Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, et al. Dinov2: Learning robust visual features without supervision. *arXiv preprint arXiv:2304.07193*, 2023. 1, 2, 3

[41] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al. Pytorch: An imperative style, high-performance deep learning library. *Advances in neural information processing systems*, 32, 2019. 7

[42] Hieu Pham, Zihang Dai, Golnaz Ghiasi, Kenji Kawaguchi, Hanxiao Liu, Adams Wei Yu, Jiahui Yu, Yi-Ting Chen, Minh-Thang Luong, Yonghui Wu, et al. Combined scaling for zero-shot transfer learning. *Neurocomputing*, 555: 126658, 2023. 3

[43] 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. 1, 2, 3, 8

[44] Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. Exploring the limits of transfer learning with a unified text-to-text transformer. *The Journal of Machine Learning Research*, 21(1):5485–5551, 2020. 6

[45] Sujeevan Ratnasingham and Paul DN Hebert. Bold: The barcode of life data system (http://www. barcodinglife. org). *Molecular ecology notes*, 7(3):355–364, 2007. 2

[46] Joseph Redmon and Ali Farhadi. Yolov3: An incremental improvement. *arXiv preprint arXiv:1804.02767*, 2018. 8

[47] 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. 8

[48] RK Samanta and Indrajit Ghosh. Tea insect pests classification based on artificial neural networks. *International Journal of Computer Engineering Science (IJCES)*, 2(6):1–13, 2012. 2

[49] Christoph Schuhmann, Romain Beaumont, Richard Vencu, Cade Gordon, Ross Wightman, Mehdi Cherti, Theo Coombes, Aarush Katta, Clayton Mullis, Mitchell Wortsman, et al. Laion-5b: An open large-scale dataset for training next generation image-text models. *Advances in Neural Information Processing Systems*, 35:25278–25294, 2022. 2

[50] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014. 1, 8

[51] Nigel E Stork. How many species of insects and other terrestrial arthropods are there on earth? *Annual review of entomology*, 63:31–45, 2018. 2

[52] Christian Szegedy, Wei Liu, Yangqing Jia, Pierre Sermanet, Scott Reed, Dragomir Anguelov, Dumitru Erhan, Vincent Vanhoucke, and Andrew Rabinovich. Going deeper with

convolutions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1–9, 2015. 1, 8

[53] Thanh-Dat Truong, Chi Nhan Duong, Ngan Le, Son Lam Phung, Chase Rainwater, and Khoa Luu. Bimal: Bijective maximum likelihood approach to domain adaptation in semantic scene segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 8548–8557, 2021. 3

[54] Thanh-Dat Truong, Quoc-Huy Bui, Chi Nhan Duong, Han-Seok Seo, Son Lam Phung, Xin Li, and Khoa Luu. Direcformer: A directed attention in transformer approach to robust action recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 20030–20040, 2022. 3

[55] Thanh-Dat Truong, Ravi Teja Nvs Chappa, Xuan-Bac Nguyen, Ngan Le, Ashley PG Dowling, and Khoa Luu. Otadapt: Optimal transport-based approach for unsupervised domain adaptation. In *2022 26th International Conference on Pattern Recognition (ICPR)*, pages 2850–2856. IEEE, 2022. 1

[56] Thanh-Dat Truong, Chi Nhan Duong, Kha Gia Quach, Ngan Le, Tien D Bui, and Khoa Luu. Liaad: Lightweight attentive angular distillation for large-scale age-invariant face recognition. *Neurocomputing*, 543:126198, 2023. 3

[57] Thanh-Dat Truong, Ngan Le, Bhiksha Raj, Jackson Cothren, and Khoa Luu. Fredom: Fairness domain adaptation approach to semantic scene understanding. In *IEEE/CVF Computer Vision and Pattern Recognition (CVPR)*, 2023.

[58] Thanh-Dat Truong, Hoang-Quan Nguyen, Bhiksha Raj, and Khoa Luu. Fairness continual learning approach to semantic scene understanding in open-world environments. *Advances in Neural Information Processing Systems*, 36, 2024. 3

[59] 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. 2, 3

[60] Kanesh Venugoban and Amirthalingam Ramanan. Image classification of paddy field insect pests using gradient-based features. *International Journal of Machine Learning and Computing*, 4(1):1, 2014. 2

[61] Jiangning Wang, Congtian Lin, Liqiang Ji, and Aiping Liang. A new automatic identification system of insect images at the order level. *Knowledge-Based Systems*, 33:102–110, 2012. 2

[62] Peng Wang, An Yang, Rui Men, Junyang Lin, Shuai Bai, Zhikang Li, Jianxin Ma, Chang Zhou, Jingren Zhou, and Hongxia Yang. Ofa: Unifying architectures, tasks, and modalities through a simple sequence-to-sequence learning framework. In *International Conference on Machine Learning*, pages 23318–23340. PMLR, 2022. 3

[63] Rujing Wang, Liu Liu, Chengjun Xie, Po Yang, Rui Li, and Man Zhou. Agripest: A large-scale domain-specific benchmark dataset for practical agricultural pest detection in the wild. *Sensors*, 21(5):1601, 2021. 2, 3

[64] Tan Wang, Kevin Lin, Linjie Li, Chung-Ching Lin, Zhengyuan Yang, Hanwang Zhang, Zicheng Liu, and Lijuan Wang. Equivariant similarity for vision-language foundation models. *arXiv preprint arXiv:2303.14465*, 2023. 3

[65] Zirui Wang, Jiahui Yu, Adams Wei Yu, Zihang Dai, Yulia Tsvetkov, and Yuan Cao. Simvlm: Simple visual language model pretraining with weak supervision. *arXiv preprint arXiv:2108.10904*, 2021. 3

[66] 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, 2, 3, 7, 8

[67] Chengjun Xie, Jie Zhang, Rui Li, Jinyan Li, Peilin Hong, Junfeng Xia, and Peng Chen. Automatic classification for field crop insects via multiple-task sparse representation and multiple-kernel learning. *Computers and Electronics in Agriculture*, 119:123–132, 2015. 2, 3

[68] Chengjun Xie, Rujing Wang, Jie Zhang, Peng Chen, Wei Dong, Rui Li, Tianjiao Chen, and Hongbo Chen. Multi-level learning features for automatic classification of field crop pests. *Computers and Electronics in Agriculture*, 152: 233–241, 2018. 2, 3

[69] Zhenda Xie, Zheng Zhang, Yue Cao, Yutong Lin, Jianmin Bao, Zhuliang Yao, Qi Dai, and Han Hu. Simmim: A simple framework for masked image modeling. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9653–9663, 2022. 1, 3

[70] 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, 2, 3, 6, 7, 8

[71] Xiaohua Zhai, Alexander Kolesnikov, Neil Houlsby, and Lucas Beyer. Scaling vision transformers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12104–12113, 2022. 2

[72] Xiaohua Zhai, Xiao Wang, Basil Mustafa, Andreas Steiner, Daniel Keysers, Alexander Kolesnikov, and Lucas Beyer. Lit: Zero-shot transfer with locked-image text tuning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18123–18133, 2022. 3, 8

[73] Xiaohua Zhai, Basil Mustafa, Alexander Kolesnikov, and Lucas Beyer. Sigmoid loss for language image pre-training. *arXiv preprint arXiv:2303.15343*, 2023. 3

[74] Shifeng Zhang, Longyin Wen, Xiao Bian, Zhen Lei, and Stan Z Li. Single-shot refinement neural network for object detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4203–4212, 2018. 8