Phyton-
International Journal of
Experimental Botany

Tech Science Press

http://doi.org/10.32604/phyton.2026.080299

ARTICLE

AP60: A Taxonomy-Guided Benchmark Dataset for Fine-Grained Pest Recognition with Feature-Level Confusion Analysis

Xianfeng Zhou[1,2,3], Shaogang Lei[1,*], Xinfeng Li[2], Zhaojie Zhang[2], Lijiao Jin[2], Jingcheng Zhang[3] and Dongmei Chen[3,*]

[1]School of Environment and Spatial Informatics, China University of Mining and Technology, Xuzhou, China
[2]Zhejiang Zhengyuan Geomatics Co., Ltd., Huzhou, China
[3]School of Artificial Intelligence, Hangzhou Dianzi University, Hangzhou, China
*Corresponding Authors: Shaogang Lei. Email: lsgang@126.com; Dongmei Chen. Email: chendongmei@hdu.edu.cn

Received: 06 February 2026; Accepted: 09 May 2026; Published: 29 June 2026

ABSTRACT: Accurate recognition of visually similar pest species remains a major challenge in agricultural vision, given that existing datasets often lack sufficient taxonomic structure, confusable categories, and quantitative analysis of class-level visual difficulty. To address these limitations, we present AP60, a taxonomy-guided benchmark dataset for fine-grained pest recognition, comprising 62,091 images from 60 pest categories and organized according to insect taxonomy. A distinctive characteristic of AP60 is the deliberate inclusion of morphologically confusable taxa, which enables more realistic evaluation of recognition models under biologically meaningful fine-grained settings. Beyond dataset construction, we introduce a feature-level confusion analysis framework to characterize the intrinsic visual structure of AP60 from two complementary aspects: intra-class consistency and inter-class overlap. Using ResNet-34 features and cosine similarity, we quantify class-wise representation similarity and relate it to downstream recognition difficulty. Benchmark evaluations were conducted under two complementary settings. In the closed-set setting, 12 supervised models achieved an average accuracy of 85.8% and an average F1-score of 85.1%, indicating that AP60 is a challenging yet stable benchmark for standard pest recognition. In the class-disjoint few-shot setting, three representative few-shot methods were evaluated on unseen pest categories, with FLoR achieving the best accuracy of 74.4% under the 5-way 5-shot protocol. These results suggest that AP60 supports both conventional supervised classification and data-efficient recognition of unseen pest categories with limited labeled samples. Further analysis shows that higher intra-class similarity is associated with better class-level accuracy, whereas lower inter-class separability is associated with increased misclassification. Validation on two additional related pest datasets shows that the same relationships remain stable after data expansion, indicating that the proposed analysis is useful not only for performance interpretation but also for identifying classes that may benefit most from targeted dataset refinement. Overall, AP60 serves as both a benchmark dataset for fine-grained pest recognition and a data-centric resource for diagnosing feature confusion in agricultural image classification.

KEYWORDS: Agricultural pest recognition; benchmark dataset; fine-grained classification; taxonomic hierarchy; few-shot recognition; feature-level confusion analysis

1 Introduction

Crop pests remain a major constraint on agricultural production as they reduce yield and quality and often drive extensive pesticide use. In practical pest management, timely and accurate species recognition is therefore essential for targeted intervention and reduced ecological cost [1,2]. However, this task is

inherently difficult: many pest species share highly similar morphology, while intra-species variation caused by viewpoint, illumination, developmental stage, and background clutter further complicates recognition [3,4]. As a result, even experienced practitioners may struggle to distinguish closely related taxa, and the same difficulty is inherited by computer vision systems operating under field conditions [5].

Early research into image-based pest identification relied on traditional machine learning pipelines, which entailed two sequential steps: manual handcrafted feature extraction and supervised classification. Feature extraction predominantly leveraged texture, shape, color, and contour descriptors, including gray-level co-occurrence matrices (GLCM) [6], GIST descriptors [7], histograms of oriented gradients (HOG) [8], speeded-up robust features (SURF) [9], scale-invariant feature transform (SIFT) [10], and local binary patterns (LBP) [11]. Classifiers including support vector machines (SVM) [12], k-nearest neighbors (KNN) [13], and decision trees (DT) [14] were then applied to the extracted feature sets for species discrimination. Although these methods showed value in constrained scenarios, their dependence on manual feature engineering limited robustness in complex agricultural environments [15]. More recently, deep learning has substantially improved pest recognition by enabling end-to-end representation learning. Convolutional neural networks and transformer-based models have been successfully applied to pest classification, detection, few-shot learning, and data augmentation, demonstrating strong performance gains over traditional pipelines [16–20]. Recent efforts have further advanced the field by integrating multi-scale feature fusion and mixed attention mechanisms into CNN architectures, achieving notable performance gains for fine-grained pest classification on large-scale benchmark datasets [21,22]. For instance, Liu et al. [23] proposed a global contrast region localization method to isolate pest regions from background noise, which were then input to a modified AlexNet for classification; Li et al. [24] fine-tuned GoogleNet on a 10-species pest dataset and achieved a classification accuracy of 98.91%. Beyond single-model optimization, Ayan et al. [25] employed a genetic algorithm for weighted fusion of three independent models to yield state-of-the-art results across three pest datasets, while Rustia et al. [26] cascaded YOLOv3 with a dedicated classification model, attaining an F1-score of 0.9 for four key pest species. Data augmentation strategies have also been widely explored to address limited sample sizes: Lu et al. [27] used generative adversarial networks (GANs) to augment pest image datasets, improving classification accuracy to 95%, and Zhou et al. [28] trained a super-resolution GAN on paired high/low-resolution stored-grain insect images to enhance low-quality imagery prior to detection. For small-sample scenarios, Li et al. [29] developed a specialized few-shot learning framework for cotton pest classification, achieving promising performance on two independent datasets. However, most pest recognition benchmarks are still evaluated under closed-set conditions, where the training and test categories are identical. This setting does not fully reflect practical scenarios in which newly observed, region-specific, or under-sampled pest categories must be recognized from only a few labeled examples. Few-shot recognition and metric-learning-based evaluation are therefore important for assessing whether learned pest representations can transfer to unseen categories.

From a data-centric perspective, three limitations remain insufficiently addressed in existing pest image resources. First, many datasets are organized primarily around practical categories or collection convenience rather than explicit insect taxonomy, which weakens their ability to reflect biologically meaningful fine-grained relationships among closely related taxa. For instance, Wu et al. [30] grouped pests into eight super-classes based on host crop affiliation, while Liu et al. [31] and Wang et al. [32] classified pests into four subclasses according to the plant tissue damaged. Wang et al. [33] established a dataset encompassing 10 fruit tree pests, 4 cereal/cruciferous crop pests, and 5 beneficial insects. Second, most benchmark evaluations focus on closed-set classification, while few studies examine class-disjoint few-shot recognition, where models must adapt to unseen pest categories with only limited labeled samples. This

limitation is particularly relevant to agricultural pest recognition because newly emerging, region-specific, or rarely observed pest species often lack sufficient annotated images. Third, current benchmark studies mainly report classification accuracy, while offering limited quantitative analysis of why some classes are intrinsically easier or harder to recognize [34]. Initial attempts at pest dataset evaluation have relied on traditional computer vision metrics: Liu et al. [31] used color histograms and oriented FAST and rotated BRIEF (ORB) features to characterize a forest pest dataset, and Wang et al. [35] applied hash algorithms to quantify inter-species similarity in an agricultural pest dataset. In particular, the field still lacks a structured way to characterize two properties that are central to fine-grained recognition: intra-class consistency, which describes how visually coherent samples of the same class are, and inter-class overlap, which describes how strongly different classes resemble one another in feature space. This gap is especially important for pest recognition because many operational errors arise not from generic visual difficulty, but from confusion among morphologically similar species [36,37].

To address these issues, we present AP60, a taxonomy-guided benchmark dataset for fine-grained pest recognition. AP60 contains 62,091 images from 60 pest categories and is organized according to insect taxonomy to better preserve hierarchical biological relationships among classes. A distinctive feature of AP60 is that it deliberately includes visually confusable taxa, including species that are morphologically similar to major agricultural pests. This design makes the dataset more suitable for realistic fine-grained recognition research than a simple category collection. In addition, we establish benchmark results under both standard closed-set classification and class-disjoint few-shot recognition settings. Twelve representative deep learning models are evaluated for supervised pest recognition, while three few-shot methods are tested under a 5-way 5-shot setting with non-overlapping training, validation, and test categories. Beyond dataset construction, this study introduces a feature-level confusion analysis framework for dataset characterization. Rather than presenting feature confusion as a standalone methodological breakthrough, we use deep feature similarity to analyze the intrinsic visual structure of AP60 from two complementary aspects: intra-class consistency and inter-class overlap. This analysis can help explain class-level recognition difficulty, complement conventional classification confusion matrices, and provide a data-centric perspective for identifying bottleneck categories that may benefit from targeted dataset refinement. We further validate these relationships on two related pest datasets with data expansion, showing that feature-level confusion measures can provide interpretable guidance for dataset diagnosis and improvement.

The main contributions of this study are as follows: First, we construct AP60, a large-scale, taxonomy-guided benchmark dataset for fine-grained pest recognition, with explicit attention to hierarchical taxonomic structure and visually confusable categories. Second, we provide benchmark evaluations under both closed-set and few-shot settings, including 12 representative deep learning models for supervised recognition and three few-shot methods for class-disjoint 5-way 5-shot pest recognition. Third, we introduce a feature-level confusion analysis framework that links deep feature similarity with class-wise recognition difficulty, thereby offering an interpretable and data-centric tool for analyzing and improving fine-grained pest datasets. The remainder of this paper is organized as follows. Section 2 describes the construction of AP60, the benchmark setting, and the feature-level confusion analysis procedure. Section 3 presents the experimental results and discusses the relationships between feature confusion and model performance. Section 4 discusses the implications, limitations, and future directions of the study, and Section 5 concludes the paper.

2 Materials and Methods

2.1 AP60 Dataset Construction and Organization

AP60 is a taxonomy-guided benchmark dataset for fine-grained pest recognition, comprising 62,091 images from 60 pest categories. The dataset was designed to support visually challenging classification scenarios by organizing pest categories according to insect taxonomy and by deliberately including morphologically confusable taxa. Unlike a simple category collection, AP60 emphasizes biologically meaningful class relationships and provides a structured benchmark for evaluating fine-grained pest recognition. The final dataset contains 20 families and 60 species, with Lepidoptera representing the dominant order. For model development and evaluation, AP60 was stratified into training, validation, and test subsets at a fixed ratio of 7:2:1, resulting in 43,402, 12,452, and 6237 images, respectively. Images were collected from two public online sources: the Baidu search engine and the Global Biodiversity Information Facility (GBIF). Image collection and curation were finalized in 2024. For each target pest category, images were retrieved using a combination of scientific Latin names, Chinese common names, English common names, host crop names, pest-related terms, and life-stage-related terms. Typical search terms included the species name, common name, “pest”, “agricultural pest”, “adult”, “larva”, and crop-specific keywords such as rice, wheat, maize, cotton, and apple. For GBIF records, only entries with available image media were considered, and the associated occurrence metadata, including country, coordinates, collection record, and license information, were retained when available. Licensing and copyright traceability were considered during dataset construction. GBIF images were retained only when the license information permitted academic reuse, including CC0, CC BY, CC BY-SA, or CC BY-NC licenses. For Baidu-derived images, Baidu was used only as an image retrieval portal, and the original source URLs and access dates were recorded for traceability. Images with explicit copyright restrictions, visible commercial watermarks, logos, or unclear ownership information were excluded from the released dataset. A license and attribution file will be provided with the dataset to document image sources and reuse conditions.

A multi-stage quality control procedure was applied before final inclusion in AP60. First, images with severe blur, overexposure, underexposure, extremely low resolution, heavy occlusion, irrelevant content, or non-target organisms were removed. Second, exact duplicates were detected using file-level hash values, and near-duplicate images were screened using perceptual hashing. Image pairs with a perceptual hash Hamming distance of 5 or lower were flagged as potential duplicates and then manually verified. Duplicate removal was performed both within each category and across categories before the training, validation, and test split to reduce the risk of data leakage. All retained images were then subjected to expert taxonomic verification. Each image was independently checked by two agricultural entomology or plant protection experts according to species-level diagnostic traits and taxonomic references. When the two experts disagreed, the sample was reviewed by a third senior expert, and the final label was determined by consensus. Images that could not be confidently identified at the species level were discarded. For GBIF images, geographic information was retained from the original occurrence records. For Baidu-derived images, geographic information was recorded only when it was explicitly available from the original source; otherwise, it was marked as unavailable. Representative samples of each pest category are shown in Fig. 1, and the complete taxonomic organization of AP60 is presented in Fig. 2. Table 1 summarizes the hierarchical structure of AP60 at the order level and reports the imbalance ratio (IR) of each taxonomic group. AP60 has an overall IR of 16.9, which is substantially lower than the IR of 80.8 reported for the widely used IP102 dataset [30], indicating a relatively more balanced distribution across categories. Compared with existing pest datasets (Table 2), AP60 does not aim to be the largest resource in terms of either image

number or category number. Instead, its main value lies in three aspects: explicit taxonomic organization, deliberate inclusion of visually confusable pest-related taxa, and a relatively balanced category distribution. These characteristics make AP60 suitable for analyzing how taxonomic structure and feature-level visual similarity influence fine-grained pest recognition.

Figure 1: Representative image samples of each pest in the AP60 dataset.

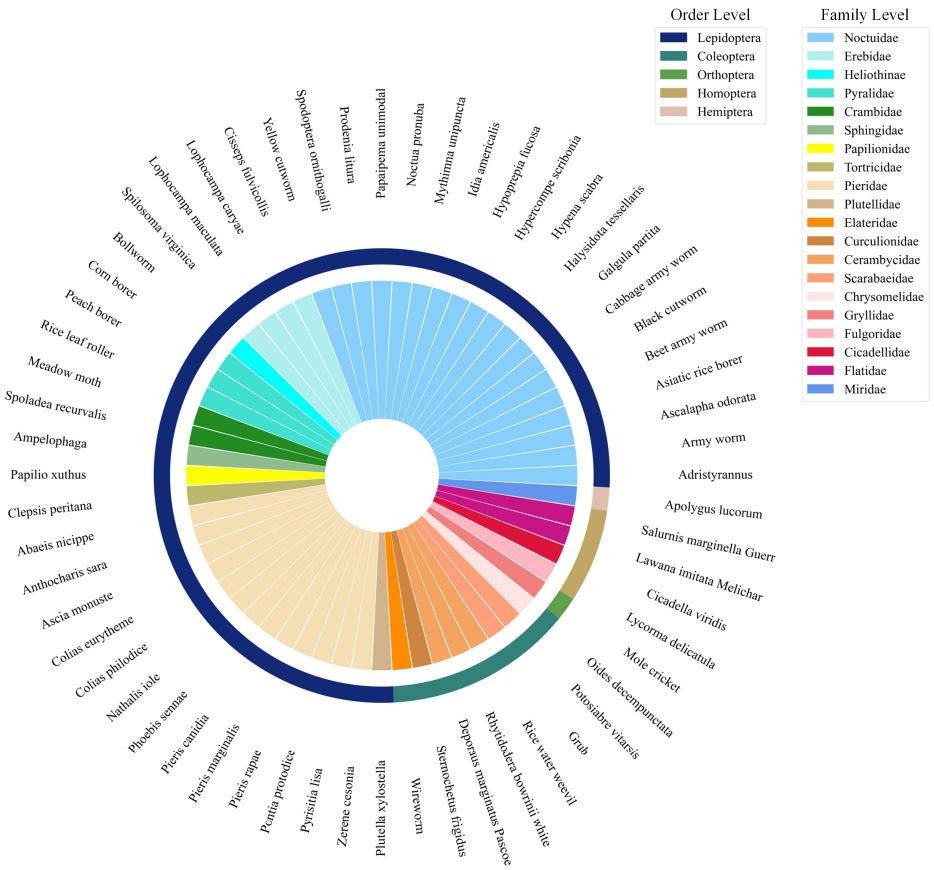

Figure 2: Taxonomic hierarchy of the AP60 dataset. The outer ring illustrates the taxonomic order of each pest species, while the inner ring corresponds to individual pest categories. Pest taxa belonging to the same family are distinguished by a uniform color scheme.

Table 1: The structures of AP60 at the different hierarchical levels and the imbalance ratio of each order.

Order	Families	Species	Samples	Imbalance Ratio (IR)
Lepidoptera	10	46	52,027	5.7
Coleoptera	5	8	2872	3.7
Orthoptera	1	1	1013	1
Homoptera	3	4	5445	7.3
Hemiptera	1	1	734	1
AP60	20	60	62,091	16.9

Note: Calculation of Imbalance ratio (IR) could be referred to [38].

Table 2: Comparison of AP60 with existing agricultural pest datasets.

Dataset	Classes	Samples	Image Source	Taxonomic Hierarchy	Confusable Taxa Design	Few-Shot Split	Feature-Level Analysis	Available
(Xie et al., 2015) [39]	24	1440	Outdoor	NR	NR	NR	NR	Y
(Xie et al., 2018) [40]	40	4500	Three public datasets	NR	NR	NR	NR	Y
(Deng et al., 2018) [41]	10	563	Online and outdoor	NR	NR	NR	NR	Y
IP102 (Wu et al., 2019) [30]	102	75,222	Online	NR	NR	NR	NR	Y
(Li et al., 2020) [24]	10	5629	Online and outdoor	NR	NR	NR	NR	Y
CPAF (Wang et al., 2020) [33]	20	4909	Online and outdoor	NR	NR	NR	NR	Y
(Alves et al., 2020) [42]	15	1500	Online and outdoor	NR	NR	NR	NR	N
(Kusrini et al., 2020) [5]	16	510	Outdoor	NR	NR	NR	NR	Y
(Khanramaki et al., 2021) [43]	10	1774	Outdoor	NR	NR	NR	NR	N
(Li et al., 2021) [44]	6	6000	Outdoor	NR	NR	NR	NR	N
(Yang et al., 2021) [45]	58	7344	Outdoor	NR	NR	NR	NR	N
AgrFIP20 (Cheng et al., 2021) [46]	20	2496	Outdoor	NR	NR	NR	NR	N
(Wei et al., 2022) [47]	12	4182	Online and outdoor	NR	NR	NR	NR	N
IP67 (Liu et al., 2022) [31]	67	67,953	Online	NR	NR	NR	NR	N
IP41 (Wang et al., 2022) [32]	41	46,615	Online	NR	NR	NR	NR	Y
AP60	60	62,091	Online	Y	Y	Y	Y	Y

Note: 'Classes' refers to the number of pest categories. 'Samples' denotes the total number of images in each dataset. 'Available' indicates whether the dataset is publicly accessible. 'Y' stands for 'Yes', 'N' stands for 'No', and 'NR' indicates that the corresponding property was not explicitly reported in the original dataset publication.

2.2 Evaluation Framework

To evaluate AP60 from both a recognition and a data-centric perspective, we employed two complementary components: benchmark classification experiments and feature-level confusion analysis (Fig. 3). The benchmark experiments quantify predictive performance across multiple deep architectures, whereas the feature-level analysis characterizes the intrinsic visual structure of the dataset in representation space. This separation is important because conventional classification metrics describe model outputs, while feature-level similarity provides additional insight into why some categories are consistently easy or difficult to distinguish. In the revised manuscript, we therefore present feature confusion analysis not as a standalone methodological breakthrough, but as an interpretable analytical framework for dataset characterization and diagnosis.

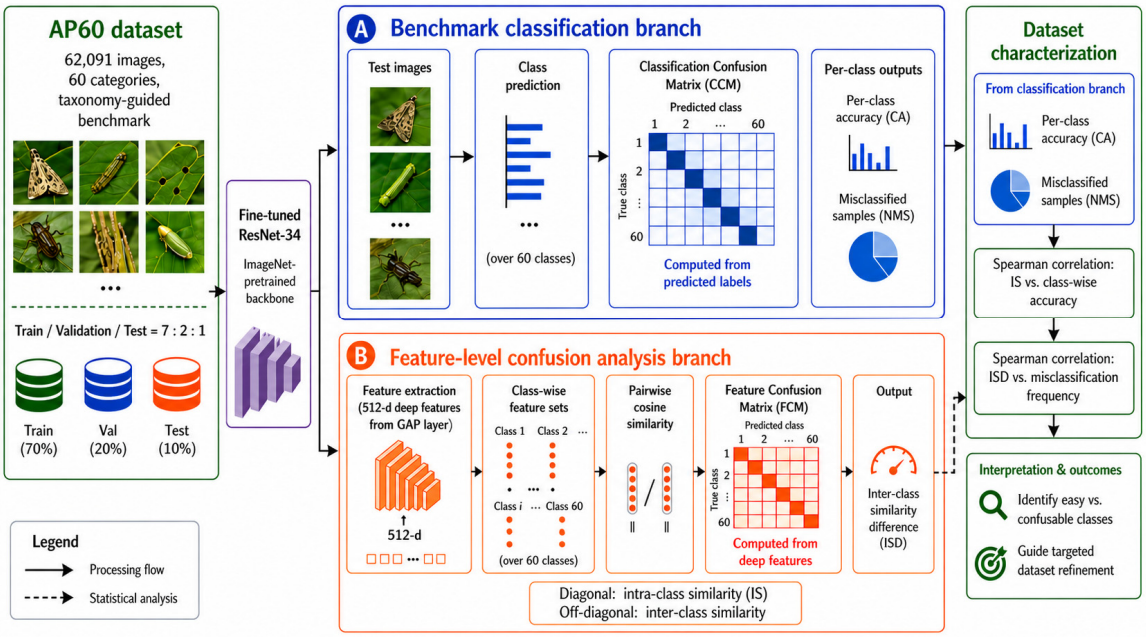

Figure 3: Workflow of benchmark classification, feature-level confusion analysis, and data characterization for AP60. A: The blue branch shows classification evaluation based on predicted labels, B: whereas the orange branch shows feature-level confusion analysis based on deep features. Solid arrows indicate processing flow, and dashed arrows indicate statistical association analysis.

2.2.1 Feature-Level Confusion Analysis

Fine-grained pest recognition is strongly affected by two dataset properties: intra-class consistency, which reflects how similar samples of the same category are to one another, and inter-class overlap, which reflects how strongly different categories resemble one another in feature space. To quantify these properties, we constructed a feature confusion matrix (FCM) based on deep representations extracted from a fine-tuned ResNet-34 model. ResNet-34 was selected because it provides a practical trade-off between representation quality, computational cost, and comparability with common visual recognition baselines. For each image, a 512-dimensional feature vector was extracted from the global average pooling layer and used as the unified feature descriptor.

Pairwise similarity between feature vectors was computed using cosine similarity, which is less sensitive to feature magnitude than Euclidean distance and is therefore more suitable for comparing directional consistency in high-dimensional representation space. The resulting feature confusion matrix contains two types of information. The diagonal entries represent intra-class similarity (IS), calculated as the mean pairwise similarity among samples within the same pest category. The off-diagonal entries represent inter-class similarity (ICS), calculated as the mean pairwise similarity between samples from two different pest categories. In this study, the FCM is used to summarize the visual structure of AP60 rather than to replace conventional classification evaluation. To link feature structure with recognition performance, we analyzed two relationships. First, the correlation between IS and class-wise classification accuracy was evaluated using Spearman's rank correlation coefficient, which is robust to non-normality and outliers. Second, inter-class distinguishability was summarized using the inter-class similarity difference (ISD), where smaller ISD values indicate stronger overlap between a target class and competing classes

and therefore a higher risk of misclassification. We then examined the relationship between ISD and the number of misclassified samples per class.

The cosine similarity between two feature vectors was calculated as:

$$S(f_a, f_b) = \sum_{i=1}^{n} \frac{\left(f_a^T f_b\right)}{n\|f_a\|_2\|f_b\|_2} \tag{1}$$

where f_a and f_b denote the 512-dimensional deep feature vectors of two images extracted from the GAP layer of ResNet-34. $S(f_a, f_b)$ ranges from −1 to 1, where larger values indicate more similar feature directions.

For pest category i, the intra-class similarity was calculated as:

$$IS_i = \frac{2}{n_i(n_i-1)} \sum_{a<b, y_a=y_b=i} S(f_a, f_b) \tag{2}$$

where n_i denotes the number of samples in category i.

The mean inter-class similarity between category i and category j was calculated as:

$$ICS_{ij} = \frac{1}{n_i n_j} \sum_{a\,:\,y_a=i} \sum_{b\,:\,y_b=j} S(f_a, f_b) \tag{3}$$

The inter-class similarity difference of category i was defined as:

$$ISD_i = M_{ii} - \frac{1}{K-1} \sum\nolimits_{j\neq i} M_{ij} \tag{4}$$

where K = 60 denotes the number of pest categories, M_{ii} denotes the intra-class similarity of category i, and M_{ij} denotes the inter-class similarity between category i and category j. A smaller ISD_i indicates that the target category is closer to other categories in feature space and is therefore more likely to be confused.

Spearman's rank correlation coefficient was calculated as:

$$\rho_{Spearman} = corr[rank(X),\ rank(Y)] \tag{5}$$

or, when no tied ranks are present:

$$\rho_{Spearman} = 1 - \left[6\sum\nolimits_{i=1}^{n} \mathrm{d}_i^2\right] / \left[\mathrm{n}\left(\mathrm{n}^2-1\right)\right] \tag{6}$$

where X and Y denote two class-wise variable vectors. For the IS-CA analysis, X is the vector of intra-class similarity values and Y is the vector of class-wise accuracies. For the ISD-NMS analysis, X is the vector of ISD values and Y is the vector of misclassified sample numbers. n denotes the number of pest categories included in the correlation analysis, and d_i denotes the difference between the ranks of the ith category in the two compared variables.

2.2.2 Benchmark Models and Training Protocol

To establish a benchmark for AP60, we fine-tuned 12 ImageNet-pretrained deep learning models, including representative convolutional neural networks and vision transformers: ResNet-34, ResNet-50 [48], DenseNet-121 [49], MobileNetV1 [50], MobileNetV2 [51], MobileNetV3 [52], ShuffleNetV2 [53], GhostNet [54], Res2Net50 [55], ViT-224, ViT-384, and Swin Transformer [56,57]. This model set spans lightweight, standard,

and transformer-based architectures, enabling a broad comparison of recognition performance under a unified transfer-learning setting [58]. In the current manuscript, all images were resized to 224 × 224 for most models, with an additional 384 × 384 setting used for ViT-384. Online augmentation consisted of random horizontal and vertical flipping with a probability of 0.5. The optimizer was Adam, the initial learning rate was 1×10^{-4}, the learning rate was decayed to 1×10^{-5} after 10 epochs, the batch size was 16, and the total number of training epochs was 20. For evaluation, model performance was measured using accuracy, precision, recall, F1-score, parameter count, and floating-point operations (FLOPs). In addition to species-level evaluation, we also conducted hierarchical assessment at the family and order levels by treating predictions within the same higher taxonomic rank as correct for the corresponding coarser-grained evaluation. This hierarchical protocol is retained because it is well aligned with the taxonomy-guided design of AP60 and provides a biologically meaningful perspective on classification difficulty across different levels of granularity.

3 Results

3.1 Benchmark Classification Performance on AP60

To establish baseline recognition performance on AP60, we evaluated 12 representative deep learning architectures under the unified training protocol described in Section 2.2.2. The evaluated models include standard convolutional neural networks, lightweight CNNs, multi-scale CNN architectures, and vision transformer-based models. Accuracy, precision, recall, F1-score, parameter number, and FLOPs were used to characterize both recognition performance and computational cost. The complete results are reported in Table 3. Overall, all models achieved species-level accuracies higher than 80%, indicating that the AP60 dataset can support stable supervised training across different model families. However, the performance differences among architectures also show that AP60 remains a challenging fine-grained recognition benchmark. Among the evaluated models, ViT-384 obtained the highest overall accuracy of 89.4% and the highest F1-score of 88.7%, followed by Swin Transformer with an accuracy of 88.9% and an F1-score of 88.3%. DenseNet121 achieved the third highest F1-score of 87.2%, while maintaining a much smaller parameter size and lower computational cost than the transformer-based models. In contrast, ShuffleNetV2 produced the lowest accuracy of 81.4%, although it required only 2.26 M parameters and 0.28 G FLOPs. These results suggest a clear trade-off between model capacity, computational complexity, and fine-grained recognition performance. The comparison between ViT-224 and ViT-384 further indicates that input resolution affects fine-grained pest recognition. Increasing the input size from 224 × 224 to 384 × 384 improved the accuracy from 86.3% to 89.4% and the F1-score from 86.0% to 88.7%. This improvement suggests that higher-resolution inputs can preserve more subtle morphological information, such as wing margins, body texture, and color patterns, which are important for distinguishing visually similar pest species. Nevertheless, this performance gain was accompanied by a substantial increase in FLOPs from 16.85 G to 49.35 G, implying that high-resolution transformer models may be less suitable for resource-constrained deployment scenarios.

In addition to overall species-level recognition, we evaluated the top three models ranked by F1-score at different taxonomic levels, including order, family, and species. In the hierarchical evaluation, predictions within the same family were treated as correct for family-level assessment, and predictions within the same order were treated as correct for order-level assessment. As shown in Fig. 4, the average accuracy of the three models reached 98.3% at the order level and 95.4% at the family level, but decreased to 88.6% at the species level. This progressive decline is consistent with the taxonomy-guided design of AP60: higher-level taxa are visually more separable, whereas species-level classification requires the model to distinguish subtle morphological differences among closely related or visually confusable categories. This hierarchical trend

also supports the necessity of evaluating pest recognition beyond a single overall accuracy value. A model may perform well at coarse taxonomic levels but still confuse species within the same family. Therefore, species-level accuracy provides a stricter measure of fine-grained recognition ability, while family- and order-level results help reveal how classification difficulty changes along the taxonomic hierarchy. This result is particularly relevant for AP60 because the dataset was designed to include morphologically similar taxa rather than only visually distinct pest categories.

Table 3: Precision, recall, F1-score, accuracy, parameter number, and FLOPs of the 12 benchmark models.

Model	Precision (%)	Recall (%)	F1-Score (%)	Accuracy (%)	Params (M)	FLOPs (G)
ViT384	89.0	88.4	**88.7**	89.4	86.42	49.35
Swin Transformer	88.7	87.9	**88.3**	88.9	88.00	15.40
DenseNet121	87.3	87.1	**87.2**	87.5	7.98	5.69
Res2Net50	86.9	86.2	86.5	87.6	25.70	8.52
ViT224	86.2	85.7	86.0	86.3	86.42	16.85
ResNet50	85.4	84.9	85.2	86.0	21.80	8.19
ResNet34	84.6	83.6	84.1	84.7	25.56	7.36
MobileNetV3	84.4	83.4	83.9	84.5	5.47	0.45
MobileNetV2	83.9	83.7	83.8	84.9	3.44	0.60
GhostNet	83.8	83.2	83.5	83.9	5.20	0.29
MobileNetV1	83.4	83.0	83.2	83.9	4.19	1.11
ShuffleNetV2	80.7	80.2	80.4	81.4	2.26	0.28
Average	85.4	84.8	85.1	85.8		

Note: Bold values indicate the top three models according to F1-score, which were selected for hierarchical evaluation in Fig. 4. Params and FLOPs indicate computational complexity. The average row reports only performance metrics.

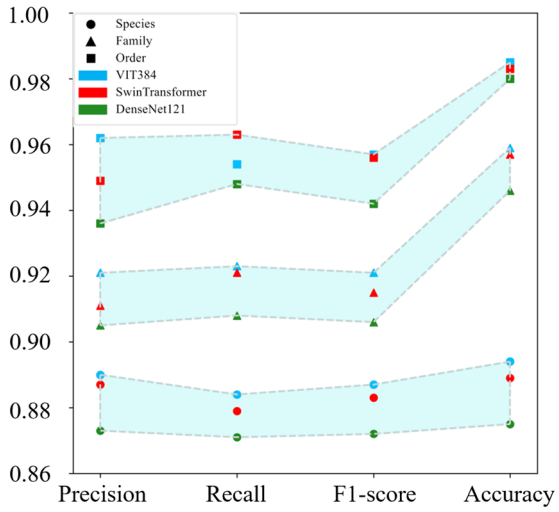

Figure 4: Hierarchical recognition performance of the top three models ranked by F1-score at the order, family, and species levels.

To further examine the spatial evidence used by CNN models during classification, Grad-CAM [59] visualization was conducted on the top three CNN-based models ranked by F1-score, namely DenseNet121, Res2Net50, and ResNet50. Transformer-based models were not included in this Grad-CAM comparison to maintain consistency in convolutional feature-map visualization. Three representative correctly classified test images were selected, corresponding to *Colias eurytheme*, *Cicadella viridis*, and *Loxostege sticticalis*. As shown in Fig. 5, the three CNN models exhibited different attention patterns. DenseNet121 generally

focused on the pest body regions, suggesting that its predictions were mainly supported by biologically relevant visual cues. By contrast, Res2Net50 and ResNet50 sometimes assigned attention to surrounding background regions, indicating that part of their classification evidence may be influenced by contextual or non-target image information. These visualization results should be interpreted as qualitative evidence rather than definitive proof of model interpretability. Nevertheless, they help illustrate an important property of AP60: fine-grained pest recognition is affected not only by model accuracy but also by whether discriminative features are extracted from the pest body itself. This observation further motivates the feature-level confusion analysis in Section 3.2, where class-level feature consistency and inter-class overlap are quantitatively examined to explain recognition difficulty more systematically.

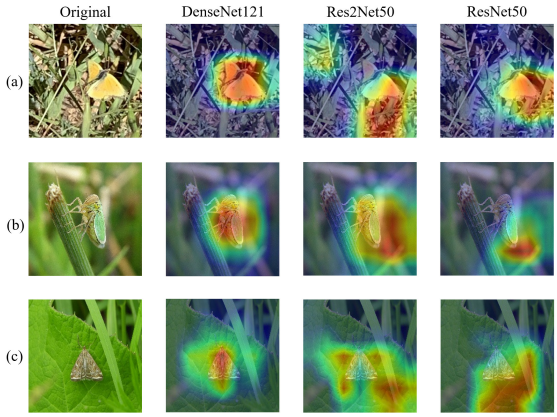

Figure 5: Grad-CAM visualizations of three representative correctly classified test images using the top three CNN-based models ranked by F1-score. Red regions indicate image areas contributing most strongly to the predicted class. (**a**) *Colias eurytheme*; (**b**) *Cicadella viridis*; (**c**) *Loxostege sticticalis*.

3.2 Feature-Level Confusion Analysis and Class-Wise Recognition Difficulty

To further examine why some pest categories are easier or more difficult to recognize, we compared the feature confusion matrix (FCM) with the conventional classification confusion matrix (CCM). These two matrices provide complementary information. The CCM summarizes the final prediction outcomes of a trained classifier, whereas the FCM describes the feature-level visual similarity structure among pest categories before the final decision layer. Therefore, the FCM is not intended to replace the CCM, but to provide an additional data-centric view for interpreting class-wise recognition difficulty. For this analysis, the fine-tuned ResNet-34 model was used because it was also adopted as the feature extractor in the feature-level confusion analysis. Deep feature vectors were extracted from the global average pooling layer for all test images, and pairwise cosine similarity was computed to construct the FCM. In Fig. 6a, the diagonal entries represent intra-class similarity (IS), which reflects feature consistency among samples from the same pest category. The off-diagonal entries represent inter-class similarity, which reflects feature overlap between different pest categories. For comparison, Fig. 6b shows the CCM derived from ResNet-34 predictions on the same test set. Both matrices were ordered according to the family-level taxonomic structure, allowing feature-level similarity patterns and prediction-level confusion patterns to be compared under the same biological hierarchy. The CCM showed that ResNet-34 achieved an overall species-level accuracy of 84.7% on AP60, but errors were unevenly distributed across taxonomic groups. Species belonging to Erebidae and Tortricidae achieved relatively high average accuracies of 93.6% and 94.1%, respectively. In contrast, species from Noctuidae and Pieridae showed lower average accuracies

of 80.8% and 82.6%, respectively. The strongest pairwise confusion occurred between *Colias eurytheme* and *Colias philodice*: 19 samples of *C. eurytheme* were misclassified as *C. philodice*, while 29 samples of *C. philodice* were misclassified as *C. eurytheme*. This result is consistent with the fine-grained nature of AP60, where visually similar species within closely related taxa are more likely to be confused.

The FCM provided a feature-level explanation for these classification patterns. Across all 60 pest categories, the mean intra-class similarity was 0.746, whereas the mean inter-class similarity was 0.465. The difference between these two values indicates that the learned representation generally preserved stronger similarity within the same pest category than between different categories. However, the matrix also revealed that inter-class similarity was not uniformly distributed. High off-diagonal similarity values were concentrated mainly within Noctuidae and Pieridae, and also appeared among several Coleoptera taxa. These regions corresponded closely to the groups with higher misclassification rates in the CCM. This suggests that the difficult classes in AP60 are not randomly distributed, but are partly associated with taxonomic and morphological similarity. The comparison between FCM and CCM highlights the different roles of the two matrices. The CCM identifies where the classifier made mistakes, but it does not directly explain whether these errors arise from intrinsic visual similarity among classes. In contrast, the FCM reveals whether different classes occupy overlapping regions in the learned feature space. For example, the Pieridae group showed strong inter-class feature similarity in the FCM, which helps explain why classification errors within this group were more frequent. Thus, the FCM provides a useful diagnostic complement to the CCM by revealing feature-level class relationships that are not fully visible from prediction counts alone.

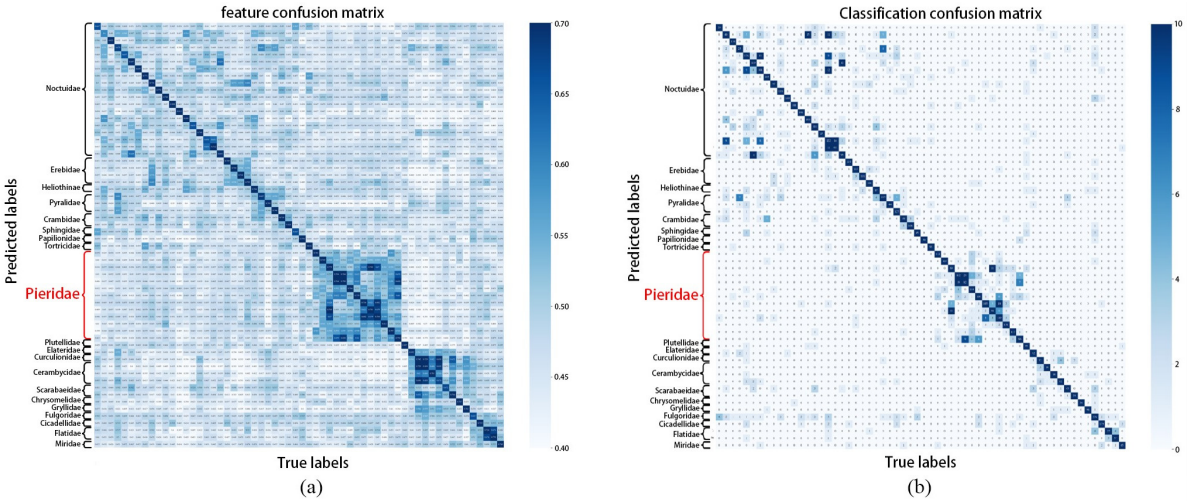

Figure 6: Feature-level and prediction-level confusion patterns of AP60. (**a**) Feature confusion matrix (FCM) constructed from ResNet-34 deep features using cosine similarity. Diagonal entries indicate intra-class similarity, and off-diagonal entries indicate inter-class similarity. (**b**) Classification confusion matrix (CCM) derived from ResNet-34 predictions on the test set. Both matrices are ordered according to family-level taxonomy to facilitate comparison between feature-level similarity and prediction-level confusion.

To quantify the relationship between feature-level structure and recognition performance, we further analyzed the association between intra-class similarity (IS) and class-wise accuracy (CA). For ResNet-34, Spearman's correlation coefficient between IS and CA was 0.611, with a p-value of 2.23×10^{-7}, indicating a significant positive relationship. This result suggests that categories with more internally consistent feature representations tend to achieve higher classification accuracy. The same analysis was extended to the other 11 benchmark models. As shown in Table 4, the correlation coefficients were consistently positive, with

an average value of 0.595 ± 0.038, indicating that the relationship between intra-class feature consistency and class-wise recognition performance was stable across model architectures. We also examined the relationship between inter-class similarity difference (ISD) and the number of misclassified samples (NMS). For ResNet-34, the Spearman correlation coefficient between ISD and NMS was -0.603, with a p-value of 3.44×10^{-7}. Because a smaller ISD indicates stronger overlap between a target category and other categories, this negative correlation means that classes with lower separability tend to produce more misclassified samples. Across all 12 models, the average correlation coefficient was -0.540 ± 0.041, further supporting the association between feature-level overlap and classification confusion. These results should be interpreted as evidence of diagnostic association, rather than as proof of a causal mechanism. The observed correlations are consistent with the expected behavior of representation learning: categories with high intra-class consistency are easier to model, whereas categories with high inter-class overlap are more difficult to separate. The value of the FCM therefore lies not in replacing established classification metrics, but in providing a quantitative way to identify which pest categories are intrinsically more ambiguous in feature space. Such information is useful for dataset diagnosis, error interpretation, and targeted data refinement, especially for fine-grained recognition tasks involving visually confusable taxa. Overall, the combined analysis of the FCM and CCM indicates that AP60 contains both well-separated and highly confusable pest categories. This supports the use of AP60 as a fine-grained benchmark dataset and demonstrates that feature-level confusion analysis can provide interpretable information beyond overall accuracy and standard confusion matrices.

Table 4: Associations between feature-level confusion metrics and class-wise recognition difficulty across benchmark models.

Model	Classification Accuracy vs. Intra-Class Similarity	Misclassification Frequency vs. ISD
ResNet34	0.611	−0.603
ResNet50	0.582	−0.489
DenseNet121	0.647	−0.553
GhostNet	0.551	−0.525
Res2Net50	0.583	−0.556
ShuffleNetV2	0.624	−0.561
Swin Transformer	0.607	−0.483
MobileNetV1	0.587	−0.551
MobileNetV2	0.545	−0.513
MobileNetV3	0.670	−0.546
ViT-224	0.545	−0.613
ViT-384	0.586	−0.485
Average	0.595 ± 0.038	-0.540 ± 0.041

Note: IS denotes intra-class similarity, CA denotes class-wise accuracy, ISD denotes inter-class similarity difference, and NMS denotes the number of misclassified samples. Positive IS–CA correlations indicate that classes with higher feature consistency tend to achieve higher accuracy. Negative ISD–NMS correlations indicate that classes with lower inter-class separability tend to produce more misclassifications.

3.3 Feature-Level Confusion Analysis for Dataset Expansion Diagnosis

To further examine whether feature-level confusion metrics can provide useful diagnostic information during dataset refinement, we conducted an additional validation experiment using two related 17-class pest datasets. This experiment was not designed to prove the universal generalizability of the feature confusion matrix across all visual domains. Instead, its purpose was to assess whether intra-class similarity and inter-class separability could reflect changes in class-wise recognition difficulty after targeted data expansion.

The first dataset, denoted as Dataset-17-Baidu, contained 17 common agricultural pest species collected from the Baidu search engine, with an average of 231 images per category. The dataset was randomly divided into training and test subsets at a ratio of 7:3. The second dataset, denoted as Dataset-17-Baidu+GBIF, was constructed by adding supplementary images collected from GBIF to the training subset of Dataset-17-Baidu. On average, 596 additional images per species were added. The test subset was kept unchanged so that the effect of training data expansion could be evaluated under a comparable testing condition. For both datasets, the same model configuration and evaluation procedure were used. The classification confusion matrix and feature confusion matrix were then constructed, and four class-wise indicators were compared: intra-class similarity (IS), classification accuracy (CA), inter-class similarity difference (ISD), and number of misclassified samples (NMS). The results are summarized in Table 5. After data expansion, IS increased for most pest categories, indicating that the model learned more consistent within-class feature representations when additional training samples were introduced. ISD also increased for most categories, suggesting improved separation between target classes and visually similar competing classes. On average, the increase in IS was 0.049, accompanied by an average improvement of 4.93 percentage points in classification accuracy. Meanwhile, the average increase in ISD was 0.034, corresponding to an average reduction of 2.35 misclassified samples per category. These results indicate that the added training images generally improved both intra-class feature consistency and inter-class feature separability. The correlation analysis further supports this interpretation. In Dataset-17-Baidu, the Spearman correlation coefficient between IS and CA was 0.872 with a p-value of 5.11×10^{-6}, indicating a significant positive association between intra-class feature consistency and classification accuracy. The correlation between ISD and NMS was -0.666 with a p-value of 3.48×10^{-3}, showing that classes with lower inter-class separability tended to produce more misclassifications. After data expansion, the IS–CA correlation increased to 0.914 with a p-value of 2.86×10^{-7}, while the ISD–NMS correlation remained negative at -0.654 with a p-value of 4.37×10^{-3}. These results show that the relationships observed in AP60 were also present in the two related validation datasets.

However, the effect of data expansion was not uniform across all categories. Although most categories showed improved IS, ISD, CA, or reduced NMS, several classes did not achieve improved accuracy after additional training images were introduced. For example, some categories showed higher feature similarity values but slightly reduced classification accuracy or increased misclassification counts. This suggests that simply increasing sample quantity does not always guarantee better recognition performance. Possible reasons include source-domain differences between Baidu and GBIF images, inconsistent image quality, background variation, or the introduction of samples with greater morphological diversity. Therefore, the proposed feature-level analysis should be interpreted as a diagnostic tool for identifying potential class-level bottlenecks, rather than as a deterministic predictor of performance improvement. A more informative pattern emerged when categories were stratified according to their baseline intra-class similarity. For categories with an initial IS greater than 0.75, data expansion resulted in a relatively modest average accuracy improvement of 2.17 percentage points. In contrast, categories with an initial IS lower than 0.75 showed a larger average accuracy improvement of 6.44 percentage points after expansion. This indicates that classes with low initial intra-class consistency may benefit more from additional training samples, because the original dataset may not sufficiently cover their appearance variation. In practical dataset construction, such classes should therefore be prioritized for targeted sample collection, annotation review, and quality control. These findings extend the role of feature-level confusion analysis from post hoc error interpretation to dataset refinement diagnosis. Specifically, IS can help identify categories with insufficient internal feature consistency, while ISD can help identify categories that remain visually close to other classes and are

therefore more prone to confusion. When used together with classification accuracy and misclassification counts, these metrics provide a more complete picture of class-wise dataset difficulty. For AP60 and related pest recognition datasets, this information can support more efficient data improvement strategies, such as collecting additional images for low-IS categories, reviewing labels for classes with abnormal ISD–NMS patterns, and prioritizing visually confusable taxa for targeted augmentation.

Overall, the two 17-class validation datasets provide additional evidence that feature-level confusion metrics are useful for analyzing how data expansion affects class-wise recognition difficulty. Nevertheless, since both datasets are related to agricultural pest recognition and were constructed using a similar collection pipeline, these results should be regarded as supporting evidence within the same application domain rather than proof of broad cross-domain generalizability. Future work should evaluate the same analysis framework on more diverse fine-grained visual datasets, different feature extractors, and stronger recognition paradigms such as metric learning and foundation-model-based representations.

Table 5: Changes in feature-level confusion metrics and class-wise recognition performance before and after data expansion.

Pest	IS (1#)	IS (2#)	CA (1#)	CA (2#)	ISD (1#)	ISD (2#)	NMS (1#)	NMS (2#)
Adristyrannus	0.641	0.715↑	78.2%	81.8%↑	0.210	0.273↑	12	10↓
Ampelophaga	0.763	0.836↑	94.9%	92.9%↓	0.242	0.291↑	5	7↑
Apolygus lucorum	0.734	0.797↑	78.4%	92.2%↑	0.211	0.250↑	11	4↓
Chilo suppressalis	0.697	0.734↑	74.4%	76.9%↑	0.206	0.235↑	10	9↓
Spodoptera exigua	0.682	0.692↑	59.6%	61.5%↑	0.197	0.234↑	21	20↓
Agrotis ipsilon	0.679	0.712↑	76.9%	82.1%↑	0.203	0.267↑	9	7↓
Cicadella viridis	0.774	0.840↑	95.6%	97.4%↑	0.215	0.252↑	12	7↓
Ostrinia furnacalis	0.734	0.782↑	82.1%	86.9%↑	0.210	0.234↑	15	11↓
Lawana imitata Melichar	0.760	0.804↑	81.8%	80.3%↓	0.230	0.251↑	12	13↑
Loxostege sticticalis	0.658	0.680↑	45.5%	66.7%↑	0.200	0.231↑	18	11↓
Oides decempunctata	0.836	0.874↑	100%	98.4%↓	0.243	0.289↑	0	1↑
Carposina sasakii	0.751	0.772↑	80.8%	84.6%↑	0.251	0.280↑	5	4↓
Pieris canidia	0.830	0.875↑	97.0%	99.2%↑	0.230	0.237↑	4	1↓
Spodoptera litura	0.691	0.742↑	73.3%	73.3%	0.208	0.235↑	20	20
Cnaphalocrocis medinalis	0.780	0.823↑	83.3%	95.8%↑	0.240	0.270↑	4	1↓
Salurnis marginella Guérin	0.721	0.816↑	84.3%	89.9%↑	0.242	0.264↑	14	9↓
Agrotis segetum	0.642	0.710↑	53.3%	63.3%↑	0.201	0.228↑	14	11↓

Note: #1 denotes the original 17-class dataset collected from Baidu. #2 denotes the expanded version in which additional GBIF images were added only to the training subset. ↑ denotes increase, while ↓ denotes decrease.

3.4 Few-Shot Recognition Evaluation on AP60

To further evaluate the applicability of AP60 beyond conventional closed-set classification, we conducted an additional few-shot recognition experiment under a class-disjoint setting. This experiment was designed to assess whether AP60 can serve as a benchmark for recognizing unseen pest categories with limited labeled samples. Specifically, the 60 pest categories were divided into 38 training classes, 10 validation classes, and 12 test classes, with no category overlap among the three subsets. The class split was generated to avoid category overlap and to preserve taxonomic diversity as much as possible. The training classes were used for model optimization, the validation classes were used for model selection and hyperparameter adjustment, and the test classes were used only for final few-shot evaluation. Three representative few-shot recognition methods were selected for comparison: FLoR [60], Dara [61], and MatchingNet [62]. FLoR aims to improve representation stability by flattening high-loss regions in the

representation space. Dara improves rapid adaptation through dual adaptive representation alignment, including prototype feature alignment and normalized distribution alignment. MatchingNet is a classic metric-learning-based few-shot method that predicts query labels by learning support-query matching relationships under episodic training. All methods were evaluated under the same 5-way 5-shot setting. During testing, 1000 random few-shot tasks were constructed from the AP60 test classes, with each task containing five classes, five support samples per class, and five query samples per class. The results are summarized in Table 6. Among the three methods, FLoR achieved the best overall performance, with an accuracy of 74.4% and an F1-score of 74.3%. It also had the lowest computational cost among the three methods, with 4.93 M parameters and 0.90 G FLOPs. Dara obtained the highest precision (74.7%) but lower recall, F1-score, and accuracy than FLoR, indicating that its adaptive alignment strategy improved prediction confidence for some categories but did not fully resolve class-level confusion under the AP60 few-shot setting. In contrast, MatchingNet achieved an accuracy of 48.7% and an F1-score of 48.5%, which were substantially lower than those of FLoR and Dara. These results indicate that AP60 remains challenging under the few-shot setting, especially when the test categories are unseen during training. The large performance gap between MatchingNet and the more recent representation adaptation methods suggests that simple support-query matching is insufficient for AP60 when pest categories are visually similar. Robust and transferable representation learning appears to be particularly important for fine-grained pest recognition with limited labeled samples. Therefore, this additional experiment extends the benchmark value of AP60 from standard closed-set classification to class-disjoint few-shot recognition, directly supporting future studies on data-efficient pest recognition.

Table 6: Few-shot recognition performance of representative methods on AP60 under the 5-way 5-shot setting.

Model	Precision (%)	Recall (%)	F1-Score (%)	Accuracy (%)	Params (M)	FLOPs (G)
FLoR	74.4	74.6	74.3	74.4	4.93	0.90
Dara	74.7	72.7	71.8	72.7	4.97	3.34
MatchingNet	48.5	48.7	48.5	48.7	0.53	6.01

Note: AP60 was split into 38 training classes, 10 validation classes, and 12 test classes without category overlap. All methods were evaluated on 1000 randomly sampled 5-way 5-shot tasks from the test classes. Each task contained five support samples and five query samples per class. Values are averaged over all test episodes. Params and FLOPs indicate model complexity. FLOPs were computed under the same episodic inference setting and therefore include support-query matching operations.

4 Discussion

This study constructed AP60, a taxonomy-guided benchmark dataset for fine-grained pest recognition, containing 62,091 images from 60 pest categories. Accurate pest recognition is an essential prerequisite for integrated pest management and targeted pesticide application, because species-level identification directly affects monitoring, intervention, and management decisions [1,2]. Unlike a simple collection of pest images, AP60 was organized according to insect taxonomy and deliberately included visually confusable taxa. This design is important because the main challenge in practical pest recognition is not only distinguishing visually distinct species, but also separating closely related or morphologically similar taxa. Therefore, the value of AP60 lies not in claiming to be the largest pest dataset, but in providing a structured benchmark for analyzing how taxonomic relationships and visual similarity influence recognition difficulty. The necessity of such a dataset can be better understood in relation to existing pest image resources. Several large-scale or task-specific pest datasets have advanced agricultural computer vision, including IP102 [30], IP67 [31], IP41 [32], CPAF [33], AP162 [34], Pest24 [35], and cross-life-stage pest benchmarks [36]. These datasets have contributed substantially to model development by increasing category coverage, sample

size, or ecological realism. However, many existing resources emphasize category expansion or detection scenarios rather than explicitly analyzing the taxonomic and feature-level sources of fine-grained confusion. AP60 complements these datasets by combining taxonomic organization, relatively balanced category distribution, and feature-level confusion analysis. This combination is particularly relevant for species-level pest recognition, where subtle morphological differences often determine classification difficulty.

The benchmark experiments showed that AP60 can support stable supervised learning across different deep architectures, while still presenting clear fine-grained classification challenges. Deep learning methods, including convolutional neural networks and vision transformers, have become dominant in pest recognition and related agricultural vision tasks because of their strong representation learning capacity [19–22]. In this study, all 12 evaluated models achieved species-level accuracies above 80%, and the average classification accuracy was 85.8%, and the average F1-score was 85.1%. The best-performing models were high-capacity architectures such as ViT-384 and Swin Transformer, which is consistent with the strong visual representation ability of transformer-based models [57,58]. However, these performance gains were accompanied by higher computational cost. In contrast, DenseNet121 achieved competitive performance with fewer parameters and lower FLOPs [50]. These results indicate that AP60 is suitable not only for comparing recognition accuracy, but also for evaluating the trade-off between model complexity and classification performance in agricultural pest recognition. The hierarchical evaluation further confirmed the fine-grained nature of AP60. The top models achieved higher accuracy at the order and family levels than at the species level, with average accuracies of 98.3%, 95.4%, and 88.6% at the order, family, and species levels, respectively. This progressive decline is biologically meaningful because species within the same family often share more similar morphological traits than taxa from different orders. Therefore, a high overall accuracy alone may not fully reflect a model's ability to resolve difficult species-level distinctions. Hierarchical evaluation provides a more informative view of model behavior by showing how recognition difficulty changes along taxonomic ranks. This is also consistent with the motivation of fine-grained pest recognition studies, where model performance is often constrained by subtle inter-class differences and large intra-class variation [21,22,29].

In addition to benchmark classification, this study introduced feature-level confusion analysis to characterize class-wise visual structure in the learned feature space. The classification confusion matrix (CCM) and feature confusion matrix (FCM) serve different but complementary purposes. The CCM reports where the classifier makes errors, while the FCM shows whether those errors are associated with high feature similarity among classes. In AP60, the mean intra-class similarity was higher than the mean inter-class similarity, suggesting that the learned representation generally preserved class-specific structure. However, high off-diagonal similarity values were concentrated in several taxonomic groups, especially among visually similar taxa in Noctuidae and Pieridae. These groups also showed relatively higher classification confusion, indicating that some errors are related to intrinsic visual overlap rather than random model failure. The correlation analysis further supported this interpretation. Across the benchmark models, intra-class similarity was positively associated with class-wise accuracy, whereas lower inter-class separability was associated with a larger number of misclassified samples. These relationships are consistent with expected behavior in representation learning: classes with more internally consistent features are easier to learn, whereas classes that overlap with other categories are more difficult to separate. Therefore, the main contribution of feature-level confusion analysis should be understood as diagnostic rather than theoretical. It does not replace classification accuracy, nor does it establish a new causal theory of recognition error. Instead, it provides an interpretable way to identify classes that are visually heterogeneous or strongly

confusable with other categories. In this sense, FCM complements conventional evaluation tools and visual explanation methods such as Grad-CAM by linking class-wise feature structure with recognition difficulty.

The additional 17-class dataset expansion experiment further illustrates the practical value of this diagnostic perspective. Data quality and data quantity are known to strongly influence deep learning performance in agricultural image recognition tasks [44,63–65]. In our experiment, adding GBIF images to the training set increased intra-class similarity and inter-class separability for most categories, accompanied by improved classification performance and reduced misclassification counts. More importantly, categories with lower initial intra-class similarity benefited more from data expansion than those with already high intra-class consistency. This suggests that feature-level confusion metrics can help prioritize data refinement. For example, categories with low IS may require more diverse and representative training samples, while categories with low ISD may require stricter annotation review, targeted augmentation, or discriminative feature learning. This is consistent with previous studies showing that data augmentation, few-shot learning, and data quality optimization can improve pest recognition under limited or imbalanced training conditions [27–29,44]. The newly added few-shot experiment further broadens the benchmark value of AP60. Unlike the standard closed-set evaluation, the class-disjoint 5-way 5-shot setting evaluates whether models can transfer knowledge from known pest categories to unseen categories with only a few labeled support samples. The results showed that FLoR and Dara substantially outperformed MatchingNet, suggesting that representation robustness and adaptive alignment are more effective than simple support-query matching for AP60. This finding is consistent with the feature-level confusion analysis: when visually similar pest categories occupy overlapping feature regions, robust representation adaptation becomes essential for reliable recognition under limited supervision. Therefore, AP60 can support not only conventional supervised pest classification but also data-efficient fine-grained recognition studies.

Nevertheless, these findings should be interpreted with caution. First, AP60 is primarily a web-collected image dataset. Although images were filtered and annotated by experts, web data may still contain source bias, background regularities, uneven image quality, or geographic imbalance. These factors may influence both model performance and feature similarity patterns. Second, the benchmark experiments were designed as a unified baseline rather than an exhaustive search for optimal model performance. Additional training strategies, stronger augmentation, repeated runs, cross-validation, and hyperparameter optimization would provide more robust estimates of model capability. Third, the FCM in this study was constructed using ResNet-34 features and cosine similarity. Although this setting is practical and interpretable, future studies should examine whether similar conclusions hold when using different feature extractors, foundation models, metric learning representations, or self-supervised models. Recent insect- or plant-disease-oriented pretraining studies suggest that domain-specific representation learning may provide stronger feature spaces for downstream agricultural vision tasks [3,63]. Finally, the two additional 17-class datasets were related to the same pest recognition domain and were constructed using a similar data collection pipeline. Therefore, they provide supporting evidence within the same application domain, but do not prove broad cross-domain generalizability. Future work should proceed in several directions. First, AP60 should be further expanded with more diverse field images, life stages, imaging conditions, and geographic sources, while ensuring transparent licensing and reproducible data release. Second, stronger benchmark protocols should be added, including repeated experiments, statistical significance analysis, metric learning baselines, foundation-model-based features, few-shot learning, and open-set recognition. Third, feature-level confusion analysis could be combined with active learning or targeted data augmentation to automatically identify categories that require additional sampling or expert review.

Finally, extending this analysis to other fine-grained agricultural vision tasks would help clarify the broader utility and limitations of feature-level dataset characterization.

Overall, AP60 and the proposed feature-level confusion analysis provide a data-centric perspective for fine-grained pest recognition. Rather than focusing only on model accuracy, this study highlights how taxonomic structure, intra-class consistency, and inter-class overlap jointly shape recognition difficulty. This perspective can support more transparent benchmark construction and more targeted dataset improvement for agricultural image classification.

5 Conclusion

This study presented AP60, a taxonomy-guided benchmark dataset for fine-grained pest recognition, containing 62,091 images across 60 pest categories. The dataset was organized according to insect taxonomy and included visually confusable taxa, making it suitable for evaluating recognition models under biologically meaningful fine-grained conditions. Benchmark experiments using 12 deep learning models and additional few-shot recognition methods showed that AP60 supports both closed-set supervised classification and class-disjoint data-efficient recognition. The hierarchical evaluation further demonstrated that recognition accuracy decreases from order to family and species levels, reflecting the increasing difficulty of distinguishing closely related pest taxa. Beyond dataset construction and model benchmarking, this study introduced a feature-level confusion analysis framework to characterize intra-class consistency and inter-class overlap in representation space. The comparison between the feature confusion matrix and the classification confusion matrix showed that feature-level similarity provides useful diagnostic information for interpreting class-wise recognition difficulty. Categories with higher intra-class similarity tended to achieve higher classification accuracy, whereas categories with lower inter-class separability were more likely to produce misclassifications. These findings indicate that feature-level confusion analysis can complement conventional performance metrics by identifying easy, difficult, and visually confusable classes. The dataset expansion experiment using two related 17-class pest datasets further showed that feature-level confusion metrics can help diagnose the effect of data refinement. Classes with low initial intra-class similarity benefited more from additional training samples, suggesting that these metrics may guide targeted data collection and annotation review. The few-shot evaluation further indicates that AP60 is suitable for studying transferable representation learning under limited labeled samples, which is important for practical pest recognition scenarios involving newly observed or under-sampled categories. However, the proposed analysis should be regarded as a data-centric diagnostic tool rather than a standalone methodological breakthrough. Its broader applicability should be further evaluated using more diverse datasets, stronger feature extractors, repeated experiments, and more advanced fine-grained recognition paradigms.

In summary, AP60 provides a structured benchmark for fine-grained pest recognition, and the accompanying feature-level confusion analysis offers an interpretable way to examine class-wise dataset difficulty. The study contributes a practical dataset resource and an analysis perspective for improving agricultural pest recognition through better dataset understanding and targeted data refinement.

Acknowledgement: The authors would like to thank those who helped for data collection and annotation.

Funding Statement: This study was financially supported by Zhejiang Provincial Natural Science Foundation of China (ZCLZ24F0201), National Natural Science Foundation of China (Project Nos.: 41901268, 62276086), and National Key R&D Program of China (2022YFD2000100).

Author Contributions: The authors confirm contribution to the paper as follows: study conception and design: Xianfeng Zhou and Shaogang Lei; data collection: Xianfeng Zhou and Xinfeng Li; analysis and interpretation of

results: Xianfeng Zhou, Zhaojie Zhang and Lijiao Jin; draft manuscript preparation: Xianfeng Zhou, Dongmei Chen and Jingcheng Zhang. All authors reviewed and approved the final version of the manuscript.

Availability of Data and Materials: The AP60 dataset is publicly available at: https://pan.baidu.com/s/17VmNs7rPeEJgaoT5NuX5ug?pwd=gtjn. The released package includes image files, class labels, taxonomic metadata, predefined training/validation/test split files, and source/attribution information. For long-term accessibility and reproducibility, the dataset will also be deposited on Zenodo or another persistent repository upon acceptance.

Ethics Approval: Not applicable.

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Pimentel D. Integrated Pest Management Vol. 1. In: Integrated pest management: Innovation-development process. Dordrecht, The Netherlands: Springer; 2009. p. 83–7. [CrossRef].
2. Guo B, Wang J, Guo M, Chen M, Chen Y, Miao Y. Overview of pest detection and recognition algorithms. Electronics. 2024;13(15):3008. [CrossRef].
3. Nguyen HQ, Truong TD, Nguyen XB, Dowling A, Li X, Luu K. 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 2024; 2024 Jun 16–22; Seattle, WA, USA. p. 21945–55. [CrossRef]. doi:10.1109/cvpr52733.2024.02072.
4. Venkateswara SM, Padmanabhan J. Deep learning based agricultural pest monitoring and classification. Sci Rep. 2025;15:8684. [CrossRef].
5. Kusrini K, Suputa S, Setyanto A, Agastya IMA, Priantoro H, Chandramouli K, et al. Data augmentation for automated pest classification in Mango farms. Comput Electron Agric. 2020;179:105842. [CrossRef].
6. Zhu LQ, Zhang Z. Auto-classification of insect images based on color histogram and GLCM. In: Proceedings of the 2010 Seventh International Conference on Fuzzy Systems and Knowledge Discovery; 2010 Aug 10–12; Yantai, China. p. 2589–93. [CrossRef].
7. Mostajer Kheirkhah F, Asghari H. Plant leaf classification using GIST texture features. IET Comput Vis. 2019;13(4):369–75. [CrossRef].
8. Venugoban K, Ramanan A. Image classification of paddy field insect pests using gradient-based features. Int J Mach Learn Comput. 2014;4(1):1. [CrossRef].
9. Pattnaik G, Parvathy K. Machine learning-based approaches for tomato pest classification. Telkomnika. 2022;20(2):321. [CrossRef].
10. Solis-Sánchez LO, Castañeda-Miranda R, García-Escalante JJ, Torres-Pacheco I, Guevara-González RG, Castañeda-Miranda CL, et al. Scale invariant feature approach for insect monitoring. Comput Electron Agric. 2011;75(1):92–9. [CrossRef].
11. Kanungo P, Ghanem S, Kumari S, Naaz R, Nayak RP. LBP feature based pest identification in rice crop. Adv Manag Technol. 2020;1(1):30–5. [CrossRef].
12. Xiao D, Feng J, Lin T, Pang C, Ye Y. Classification and recognition scheme for vegetable pests based on the BOF-SVM model. Int J Agric Biol Eng. 2018;11(3):190–6. [CrossRef].
13. Vinodhkumar B, Ravi A, Sivakumar S. A framework for insect detection and classification in paddy images. In: Proceedings of the 2024 13th International Conference on System Modeling & Advancement in Research Trends (SMART); 2024 Dec 6–7; Moradabad, India. p. 34–8. [CrossRef].
14. Wen C, Guyer D. Image-based orchard insect automated identification and classification method. Comput Electron Agric. 2012;89:110–5. [CrossRef].
15. Li W, Wang D, Li M, Gao Y, Wu J, Yang X. Field detection of tiny pests from sticky trap images using deep learning in agricultural greenhouse. Comput Electron Agric. 2021;183:106048. [CrossRef].
16. Krizhevsky A, Sutskever I, Hinton GE. ImageNet classification with deep convolutional neural networks. Commun ACM. 2017;60(6):84–90. [CrossRef].

17. Chen J, Liu Q, Gao L. Deep convolutional neural networks for tea tree pest recognition and diagnosis. Symmetry. 2021;13(11):2140. [CrossRef].
18. Li L, Zhang S, Wang B. Apple leaf disease identification with a small and imbalanced dataset based on lightweight convolutional networks. Sensors. 2022;22(1):173. [CrossRef].
19. Wang X, Xiao Z, Deng Z. Swin Attention Augmented Residual Network: A fine-grained pest image recognition method. Front Plant Sci. 2025;16:1619551. [CrossRef].
20. Fang M, Tan Z, Tang Y, Chen W, Huang H, Dananjayan S, et al. Pest-ConFormer: A hybrid CNN-Transformer architecture for large-scale multi-class crop pest recognition. Expert Syst Appl. 2024;255:124833. [CrossRef].
21. Yuan Y, Sun J, Zhang Q. An enhanced deep learning model for effective crop pest and disease detection. J Imaging. 2024;10(11):279. [CrossRef].
22. Qian Y, Xiao Z, Deng Z. Fine-grained crop pest classification based on multi-scale feature fusion and mixed attention mechanisms. Front Plant Sci. 2025;16:1500571. [CrossRef].
23. Liu Z, Gao J, Yang G, Zhang H, He Y. Localization and classification of paddy field pests using a saliency map and deep convolutional neural network. Sci Rep. 2016;6:20410. [CrossRef].
24. Li Y, Wang H, Dang LM, Sadeghi-Niaraki A, Moon H. Crop pest recognition in natural scenes using convolutional neural networks. Comput Electron Agric. 2020;169:105174. [CrossRef].
25. Ayan E, Erbay H, Varçın F. Crop pest classification with a genetic algorithm-based weighted ensemble of deep convolutional neural networks. Comput Electron Agric. 2020;179:105809. [CrossRef].
26. Rustia DJA, Chao JJ, Chiu LY, Wu YF, Chung JY, Hsu JC, et al. Automatic greenhouse insect pest detection and recognition based on a cascaded deep learning classification method. J Appl Entomol. 2021;145(3):206–22. [CrossRef].
27. Lu CY, Arcega Rustia DJ, Lin TT. Generative adversarial network based image augmentation for insect pest classification enhancement. IFAC PapersOnLine. 2019;52(30):1–5. [CrossRef].
28. Zhou H, Miao H, Li J, Jian F, Jayas DS. A low-resolution image restoration classifier network to identify stored-grain insects from images of sticky boards. Comput Electron Agric. 2019;162:593–601. [CrossRef].
29. Li Y, Yang J. Few-shot cotton pest recognition and terminal realization. Comput Electron Agric. 2020;169:105240. [CrossRef].
30. Wu X, Zhan C, Lai YK, Cheng MM, Yang J. IP102: A large-scale benchmark dataset for insect pest recognition. In: Proceedings of the 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR); 2019 Jun 15–20; Long Beach, CA, USA. p. 8787–96. [CrossRef].
31. Liu Y, Liu S, Xu J, Kong X, Xie L, Chen K, et al. Forest pest identification based on a new dataset and convolutional neural network model with enhancement strategy. Comput Electron Agric. 2022;192:106625. [CrossRef].
32. Wang K, Chen K, Du H, Liu S, Xu J, Zhao J, et al. New image dataset and new negative sample judgment method for crop pest recognition based on deep learning models. Ecol Inform. 2022;69:101620. [CrossRef].
33. Wang J, Li Y, Feng H, Ren L, Du X, Wu J. Common pests image recognition based on deep convolutional neural network. Comput Electron Agric. 2020;179:105834. [CrossRef].
34. Wang K, Li W, Wu X, Xu J, Wang ZW, Yang S. AP162: A large-scale dataset for agricultural pest recognition. Comput Electron Agric. 2025;237:110520. [CrossRef].
35. Wang QJ, Zhang SY, Dong SF, Zhang GC, Yang J, Li R, et al. Pest24: A large-scale very small object data set of agricultural pests for multi-target detection. Comput Electron Agric. 2020;175:105585. [CrossRef].
36. Han Y, Zhang C, Zhan X, Huang Q, Wang Z. Crossing multiple life stages: Fine-grained classification of agricultural pests. Plant Meth. 2024;20(1):191. [CrossRef].
37. Zhang J, Liu Z, Yu K. MSFNet-CPD: Multi-scale cross-modal fusion network for crop pest detection. arXiv:2505.02441. 2025. [CrossRef].
38. Zhu R, Guo Y, Xue JH. Adjusting the imbalance ratio by the dimensionality of imbalanced data. Pattern Recognit Lett. 2020;133:217–23. [CrossRef].
39. Xie C, Zhang J, Li R, Li J, Hong P, Xia J, et al. Automatic classification for field crop insects via multiple-task sparse representation and multiple-kernel learning. Comput Electron Agric. 2015;119:123–32. [CrossRef].
40. Xie C, Wang R, Zhang J, Chen P, Dong W, Li R, et al. Multi-level learning features for automatic classification of field crop pests. Comput Electron Agric. 2018;152:233–41. [CrossRef].

41. Deng L, Wang Y, Han Z, Yu R. Research on insect pest image detection and recognition based on bio-inspired methods. Biosyst Eng. 2018;169:139–48. [CrossRef].
42. Alves AN, Souza WSR, Borges DL. Cotton pests classification in field-based images using deep residual networks. Comput Electron Agric. 2020;174:105488. [CrossRef].
43. Khanramaki M, Askari Asli-Ardeh E, Kozegar E. *Citrus* pests classification using an ensemble of deep learning models. Comput Electron Agric. 2021;186:106192. [CrossRef].
44. Li Y, Chao X. Toward sustainability: Trade-off between data quality and quantity in crop pest recognition. Front Plant Sci. 2021;12:811241. [CrossRef].
45. Yang Z, Yang X, Li M, Li W. Small-sample learning with salient-region detection and center neighbor loss for insect recognition in real-world complex scenarios. Comput Electron Agric. 2021;185:106122. [CrossRef].
46. Cheng Z, Xia W. Fine-grained image classification on agricultural pest larvae. IOP Conf Ser Earth Environ Sci. 2021;792(1):012037. [CrossRef].
47. Wei D, Chen J, Luo T, Long T, Wang H. Classification of crop pests based on multi-scale feature fusion. Comput Electron Agric. 2022;194:106736. [CrossRef].
48. He K, Zhang X, Ren S, Sun J. Deep residual learning for image recognition. In: Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR); 2016 Jun 27–30; Las Vegas, NV, USA. p. 770–8. [CrossRef].
49. Huang G, Liu Z, Van Der Maaten L, Weinberger KQ. Densely connected convolutional networks. In: Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR); 2017 Jul 21–26; Honolulu, HI, USA. p. 4700–8. [CrossRef].
50. Howard AG, Zhu M, Chen B, Kalenichenko D, Wang W, Weyand T, et al. Mobilenets: Efficient convolu-tional neural networks for mobile vision applications. arXiv:1704.04861. 2017.
51. Sandler M, Howard A, Zhu M, Zhmoginov A, Chen LC. MobileNetV2: Inverted residuals and linear bottlenecks. In: Proceedings of the 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition; 2018 Jun 18–23; Salt Lake City, UT, USA. p. 4510–20. [CrossRef].
52. Howard A, Sandler M, Chen B, Wang W, Chen LC, Tan M, et al. Searching for MobileNetV3. In: Proceedings of the 2019 IEEE/CVF International Conference on Computer Vision (ICCV); 2019 Oct 27–Nov 2; Seoul, Republic of Korea. p. 1314–24. [CrossRef].
53. Ma N, Zhang X, Zheng HT, Sun J. ShuffleNet V2: Practical guidelines for efficient CNN architecture design. In: Proceedings of the Computer Vision—ECCV 2018; 2018 Sep 8–14; Munich, Germany. p. 122–38. [CrossRef].
54. Han K, Wang Y, Tian Q, Guo J, Xu C, Xu C. GhostNet: More features from cheap operations. In: Proceedings of the 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR); 2020 Jun 13–19; Seattle, WA, USA. p. 1580–9. [CrossRef].
55. Gao SH, Cheng MM, Zhao K, Zhang XY, Yang MH, Torr P. Res2Net: A new multi-scale backbone architecture. IEEE Trans Pattern Anal Mach Intell. 2021;43(2):652–62. [CrossRef].
56. Dosovitskiy A, Beyer L, Kolesnikov A, Weissenborn D, Zhai X, Unterthiner T, et al. An image is worth 16x16 words: Transformers for image recognition at scale. arXiv:2010.11929. 2020.
57. Liu Z, Lin Y, Cao Y, Hu H, Wei Y, Zhang Z, et al. Swin transformer: Hierarchical vision transformer using shifted windows. In: Proceedings of the 2021 IEEE/CVF International Conference on Computer Vision (ICCV); 2021 Oct 10–17; Montreal, QC, Canada. p. 10012–22. [CrossRef].
58. Deng J, Dong W, Socher R, Li LJ, Li K, Li FF. ImageNet: A large-scale hierarchical image database. In: Proceedings of the 2009 IEEE Conference on Computer Vision and Pattern Recognition; 2009 Jun 20–25; Miami, FL, USA. p. 248–55. [CrossRef].
59. Selvaraju RR, Cogswell M, Das A, Vedantam R, Parikh D, Batra D. Grad-CAM: Visual explanations from deep networks via gradient-based localization. In: Proceedings of the 2017 IEEE International Conference on Computer Vision (ICCV); 2017 Oct 22–29; Venice, Italy. p. 618–26. [CrossRef].
60. Zou Y, Liu Y, Hu Y, Li Y, Li R. Flatten long-range loss landscapes for cross-domain few-shot learning. In: Proceedings of the 2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR); 2024 Jun 16–22; Seattle, WA, USA. p. 23575–84. [CrossRef].
61. Zhao Y, Zhang T, Li J, Tian Y. Dual adaptive representation alignment for cross-domain few-shot learning. IEEE Trans PAMI. 2023;45(10):11720–32. [CrossRef].

62. Vinyals O, Blundell C, Lillicrap T, Wierstra D. Matching networks for one shot learning. arXiv:1606.04080. 2016.
63. Dong X, Wang Q, Huang Q, Ge Q, Zhao K, Wu X, et al. PDDD-PreTrain: A series of commonly used pre-trained models support image-based plant disease diagnosis. Plant Phenomics. 2023;5:54. [CrossRef].
64. Sun C, Shrivastava A, Singh S, Gupta A. Revisiting unreasonable effectiveness of data in deep learning era. In: Proceedings of the 2017 IEEE International Conference on Computer Vision (ICCV); 2017 Oct 22–29; Venice, Italy. p. 843–52. [CrossRef].
65. Barbedo JGA. Impact of dataset size and variety on the effectiveness of deep learning and transfer learning for plant disease classification. Comput Electron Agric. 2018;153:46–53. [CrossRef].