Journal of Electronic Research and Application, 2026, Volume 10, Issue 9
https://bbwpublisher.com/journal/JERA
Online ISSN: 2208-3510
Print ISSN: 2208-3502

Fine-Grained Visual Classification: From Part Discovery to Efficient Multimodal Reasoning

Tao Zhang[1], Yanli Shi[2*]

[1]College of Information and Control Engineering, Jilin University of Chemical Technology, Jilin, Jilin, China
[2]College of Science, Jilin University of Chemical Technology, Jilin, Jilin, China

Author to whom correspondence should be addressed.

Copyright: © 2026 Author(s). This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY 4.0), permitting distribution and reproduction in any medium, provided the original work is cited.

Abstract: This survey examines FGVC, in which a model must distinguish visually similar categories that nevertheless exhibit substantial within-class variation. We organize recent progress around three decisions that recur across architectures: which evidence should be selected, how local and global evidence should be integrated, and how supervision should be obtained when detailed labels are scarce. We first define the task and examine the assumptions embedded in commonly used benchmarks. We then trace the transition from convolutional part discovery and second-order feature encoding toward token-based transformer models, focusing on design principles that generalize across implementations rather than presenting a chronological catalog. The review further compares localization, feature interaction, augmentation, and parameter-efficient adaptation, and discusses how evaluation protocols affect reported gains. We conclude by examining annotation bias, long-tailed distributions, domain shift, computational constraints, and the emerging role of vision-language models. The resulting synthesis is intended to support method selection and to identify research problems that extend beyond marginal improvements on saturated benchmarks.

Keywords: Fine-grained recognition; Discriminative parts; Visual transformers; Weak supervision; Vision-language learning

Online publication: October 30, 2026

1. Introduction

Within computer vision, FGVC addresses recognition inside a visually coherent super-class, including bird species, dog breeds, vehicle variants, plant varieties, and product models. The decision boundary is seldom determined by a single object-level attribute. More often, it depends on a small configuration of cues, for example feather arrangement, muzzle geometry, wheel design, or surface texture. At the same time, viewpoint, illumination, pose, occlusion, and image quality can substantially alter the appearance of a category. FGVC is therefore a demanding test of representation learning: a successful model must suppress nuisance variation without discarding the weak signals that separate neighboring classes.

The field has developed through several overlapping phases. Early systems used object or part annotations to supervise region proposals. Weakly supervised convolutional models subsequently learned salient regions from image-level labels, while bilinear and cross-layer encoders increased the capacity to model feature interactions. Vision transformers then redirected attention from hand-crafted part pipelines toward token relations and global context. Recent systems combine selective tokens, multi-scale processing, structural priors, self-supervised pretraining, and lightweight adaptation. Although the literature is extensive, the same underlying design choices recur across these architectural changes.

This article presents a concept-driven review rather than a leaderboard. It is organized around three questions: what visual evidence is selected, how are local and global cues combined, and what assumptions are made about annotation and deployment? Methods are grouped by their role in the recognition pipeline, and limitations are discussed alongside the gains they enable. This structure makes it possible to compare methods that use different backbones but address the same information bottleneck.

2. Review scope and task definition

We use FGVC to denote single-image classification in which the labels represent fine subdivisions of a shared semantic group. Let x denote an image and y in {1,...,C} its class label. The relevant evidence is usually concentrated in an unknown collection of regions, attributes, or spatial relations. In a fully supervised setting, part boxes or keypoints may be available during training; in the more common weakly supervised setting, only the image-level label is observed. This review covers image-only methods, transformer-based models, and adaptation or multimodal techniques that directly address fine-grained recognition. Detection, segmentation, and open-set recognition are considered only when their mechanisms provide transferable insight for FGVC.

The review was assembled from the FGVC manuscripts and papers supplied with the project, with methods grouped by their learning objective and information flow. We discuss representative approaches and recurring benchmark practices rather than claiming a formal systematic meta-analysis. This qualification is important because reported performance depends on pretraining data, crop policy, image resolution, sampling strategy, and test-time augmentation. Results obtained under different protocols should therefore be interpreted as indicative comparisons, not as a definitive ranking.

3. Datasets, signals, and evaluation

Common benchmarks differ in class cardinality and in the granularity of the evidence that separates labels. As illustrated in **Table 1**, CUB-200-2011 focuses on bird species and provides part metadata; Stanford Dogs tests breed recognition under substantial pose variation; NABirds expands the label inventory while increasing visual overlap; Oxford 102 Flowers is particularly sensitive to texture and contour; and Oxford-IIIT Pet combines breed cues with large changes in pose and background. These datasets are useful but not interchangeable. A method that benefits from clean object crops may be unreliable in field photographs, while a high score on a nearly saturated flower benchmark may provide limited evidence of transfer to rare categories.

Table 1. Representative FGVC benchmarks and the cues they expose

Benchmark	Domain	Classes	Main cue	Typical risk
CUB-200-2011	Birds	200	Parts and plumage	Small objects
Stanford Dogs	Dogs	120	Face and coat	Pose / breed overlap
NABirds	Birds	555	Fine morphology	Long tail
Oxford 102 Flowers	Flowers	102	Texture and contour	Background leakage
Oxford-IIIT Pet	Pets	37	Head and fur	Intra-class pose

Top-1 accuracy remains the dominant metric, but it is insufficient for deployment-oriented comparison. Per-class recall exposes failures on rare categories, calibration indicates whether confidence reflects evidential support, and robustness tests quantify sensitivity to crops, backgrounds, and image corruption. Parameter counts and multiply-accumulate estimates are meaningful only when input resolution and hardware are reported as well. For long-tailed data, macro-averaged scores and head/middle/tail breakdowns should accompany the overall accuracy.

4. Method evolution

4.1. Convolutional localization and feature interaction

CNN-based FGVC systems generally follow one of two routes. Localization-oriented models learn a sequence of increasingly focused crops or attention maps. The classifier receives a global view together with one or more candidate parts, so the objective rewards both category evidence and informative region selection. This family is intuitive and can produce inspectable regions, yet its search procedure may amplify an early localization error. Occlusion or an unusual pose can cause the crop generator to lock onto a background patch. RA-CNN and NTS-Net are representative examples [1,2].

A second route retains a broad feature map and explicitly encodes interactions among channels, scales, or spatial locations. Bilinear pooling and related second-order operators are effective when two classes share a dominant color but differ in the co-occurrence of color and texture. Cross-layer interaction models exemplify this direction [3]. The associated cost is substantial: quadratic feature interactions and covariance-like descriptors are expensive, and their high dimensionality can be difficult to regularize when only a few images are available for each class.

Several lessons from this period remain relevant. Part discovery should be treated as uncertain evidence rather than as a hard crop; global context should be used to verify a local hypothesis; and feature diversity should not be confused with discriminative value. The same considerations reappear in modern token pruning and attention filtering, even when the backbone has changed.

4.2. Transformers and token-based reasoning

Vision transformers represent an image as an ordered set of patch representations and use attention to exchange information across distant locations [4]. This design addresses a central limitation of local convolution: relationships between spatially separated parts can be modeled without waiting for information to propagate through many layers. Global access, however, creates a new selection problem. Most patches belong to the background or to redundant portions of the object, and unrestricted attention may spread probability mass over precisely the regions that should be suppressed.

Transformer-based FGVC methods differ primarily in how they select and combine tokens. TransFG

aggregates attention to retain candidate evidence, whereas IELT treats diversity across attention heads as a voting signal [5,6]. Multi-scale variants process coarse and fine patches in parallel, and structure-aware models connect selected patches through spatial or semantic relations [7,8]. Background suppression and local concealment provide additional pressure against shortcut learning [9,10]. These mechanisms can improve both interpretability and accuracy, but their effectiveness depends strongly on the statistic used to identify informative tokens.

A useful distinction is between selection before interaction and selection after interaction. Early pruning reduces computation but risks discarding a weak cue; late reweighting preserves information but offers less acceleration. Hybrid designs retain a small global token set while allocating additional layers to a compact local set. They are particularly attractive for edge devices because expensive operations can be concentrated on the evidence most likely to affect the decision.

4.3. Weak, self-supervised, and progressive learning

The annotation burden in FGVC extends beyond drawing part boxes. Experts must agree on class boundaries, and some distinctions are meaningful only within a particular geographic or industrial context. This has motivated learning schemes that reduce the need for detailed labels. Weakly supervised models infer useful regions from image-level categories, but the signal is indirect: a classifier can minimize the loss by exploiting a correlated background or a single easy attribute. Region masking, attention regularization, and consistency across augmented views are therefore important safeguards because they encourage the model to assemble several complementary cues rather than repeatedly selecting one shortcut.

Self-supervised learning changes the source of the training signal. Contrastive objectives pull different views of the same image together while separating views from other images; masked-image objectives instead ask the encoder to recover missing content. Both can provide useful shape and texture priors before fine-grained labels are introduced. Their benefit is greatest when pretraining data resemble the target domain. Generic web images may convey broad object identity while failing to capture the subtle distinctions between neighboring cultivars or product variants. A practical strategy is to pretrain on unlabeled in-domain images and then fine-tune conservatively with class-balanced sampling.

Progressive training offers another way to manage task difficulty. A coarse stage can establish object extent or super-class structure, after which a finer stage emphasizes local patterns. Jigsaw-style and multi-granularity curricula expose the network to small regions only after it has learned a stable global representation. The ordering of objectives matters: asking a randomly initialized model to classify tiny parts often produces unstable attention, whereas a global-to-local curriculum gives the local branch a meaningful reference frame.

5. Cross-cutting design dimensions

5.1. Where is the evidence?

Localization can be explicit, implicit, or relational. Explicit methods predict boxes, points, or masks and are easiest to audit when such annotations are available. Implicit methods use class activations or attention weights as soft maps and avoid additional labels, but attention should not automatically be interpreted as an explanation. Relational methods model a set of parts together with their arrangement, which is useful when categories share individual parts but differ in geometry. A robust system should preserve uncertainty over

candidate regions and allow global context to reject a spurious local cue.

5.2. How are features combined?

The central representational trade-off is between selectivity and coverage. Pooling all spatial features provides stable coverage but dilutes rare details. Selecting only the most salient token improves focus but can miss a secondary cue. Multi-branch encoders, feature pyramids, and cross-attention make different compromises. The appropriate choice depends on object scale and the class definition: a flower may be identified from a repeated texture, whereas a bird may require both beak geometry and wing pattern.

Regularization is also a representational choice. Distillation can transfer inter-class relations from a larger teacher [11]; contrastive objectives can separate visually adjacent labels; and masked-image pretraining can provide shape priors without part boxes. These losses are complementary only when they impose distinct constraints. If their targets are redundant, the model may simply become more confident in the same shortcut.

5.3. Labels, pretraining, and adaptation

FGVC datasets are expensive to curate because class boundaries often require expert knowledge. Weak supervision lowers annotation cost but can increase susceptibility to background bias. Self-supervised pretraining improves initialization, although generic web images may underrepresent the domain-specific distinctions that matter. Parameter-efficient tuning is increasingly attractive: prompt tokens, adapters, or a small subset of weights can adapt a foundation model while preserving broad visual knowledge. The unresolved question is which parameters encode transferable visual structure and which are specific to the target taxonomy.

5.4. Augmentation, sampling, and loss design

Data augmentation in FGVC must preserve the cue that defines the label. Random crops and strong color perturbations can improve invariance, but an aggressive crop may remove a beak tip, petal edge, or logo that is genuinely diagnostic. Object-aware cropping, moderate geometric perturbation, and background replacement are safer when the target occupies only a limited share of each image. Augmentation should also reflect class frequency: rare categories benefit from diverse views, whereas head classes may require stronger regularization to prevent them from dominating the representation.

Sampling and loss design determine which errors the optimizer encounters most often. Uniform image sampling can leave tail categories underrepresented, whereas class-aware batches expose the model to difficult neighboring classes more regularly. Margin-based classification, focal weighting, and supervised contrastive losses alter the geometry of the embedding space, but none removes the need for an appropriate label hierarchy. When two labels differ by a single attribute, a hierarchical or attribute-aware target can provide a more informative gradient than a flat one-hot vector. This choice should be reported explicitly because it changes the interpretation of a benchmark comparison.

Hard-example mining is useful but can become self-reinforcing. If uncertainty arises from a mislabeled or severely occluded image, repeatedly sampling that example may teach noise. A robust pipeline should track whether difficulty persists across random seeds and augmentations, and distinguish genuine ambiguity from annotation error. Reviewing a small, high-impact subset by hand often yields more insight than adding another attention block.

6. Efficiency, reliability, and reproducibility

Accuracy improvements should be interpreted alongside the cost of region proposals, additional crops, high-resolution inputs, and multi-stage inference. A model with modest parameter growth can still be slow if it repeats attention over several views. Reporting throughput, peak memory, and latency at a stated batch size makes this trade-off visible. Fair comparisons also require the same pretraining source, augmentation policy, crop count, and test-time ensemble across baselines.

Reproducibility is a central concern in this field because seemingly minor implementation choices can have large effects. Class-balanced sampling, image resizing, label smoothing, and duplicate-image handling can all change the apparent ranking. Authors should release split files, preprocessing code, random seeds, and per-class results. Attention visualizations should be accompanied by perturbation tests or deletion curves; a heatmap alone does not establish that the highlighted region caused the prediction.

6.1. Error analysis and robustness protocols

A useful FGVC evaluation should characterize failure modes rather than report only an average score. Confusion matrices reveal whether errors are concentrated among a small group of neighboring classes. Per-class recall and head/middle/tail breakdowns show whether an improvement comes from common categories or reflects broader coverage. Calibration is equally important: a model that assigns 0.99 confidence to an uncertain species may be more costly than a slightly less accurate model that signals ambiguity. Expected calibration error, classwise reliability plots, and selective accuracy at different coverage levels are inexpensive additions to a standard report.

Robustness tests should manipulate the nuisance factors known to affect fine-grained images. Background replacement tests whether the classifier relies on scenery; blur and compression tests whether it actually uses texture; occlusion tests whether it can combine multiple parts; and cross-camera or cross-location splits probe domain shift. For transformer models, token-drop experiments can estimate how many patches are necessary. For localization models, deleting the proposed region and inserting a distractor provides a direct check that the selected part is causally useful rather than merely correlated with the output.

Error analysis can also guide annotation. When two classes are repeatedly confused, experts can determine whether the distinction is visually supported, whether the label definition is inconsistent, or whether the benchmark lacks a relevant contextual attribute. This closes the loop between model development and dataset design. A review protocol that records why an error occurred is more informative than one that records only the incorrect label.

7. Open problems

First, many benchmarks contain background or capture artifacts correlated with the labels. Models may therefore succeed without learning the intended morphology. Counterfactual backgrounds, cross-location splits, and object-level masking are practical ways to quantify this risk. Second, long-tailed categories and novel classes remain underexplored in standard supervised protocols. Tail-aware losses and prototype-based updates are promising, but they must be tested under realistic annotation budgets.

Third, domain shift is the rule rather than the exception. A field photograph, a museum image, and a mobile capture may depict the same species at different scales and with different color statistics. Test-time

adaptation and domain generalization should be evaluated without access to hidden target labels. Fourth, fine-grained recognition is moving beyond closed-set classification. Vision-language models can use descriptions of parts and attributes to support zero-shot or interactive recognition, but language priors may introduce bias and may suggest distinctions unsupported by the pixels.

Finally, efficiency and interpretability are coupled. Token pruning, quantization, and early exit reduce cost, yet an aggressive policy can remove the evidence required for a rare class. Future systems should expose a confidence-aware compute policy: allocate more computation when candidate categories remain ambiguous and exit early when multiple independent cues agree.

7.1. Vision-language and interactive recognition

Vision-language models offer a natural interface for fine-grained categories because many distinctions can be expressed as attributes: a narrow bill, a spotted wing, a short muzzle, or a particular stitching pattern. Text prompts can provide a provisional taxonomy when labeled images are scarce, and a language model can expand a class description into candidate attributes. This does not eliminate the visual problem. Descriptions may be incomplete, synonyms may be unevenly represented, and language priors may favor plausible but unsupported distinctions. A reliable system should therefore align each textual hypothesis with image evidence and expose the regions or attributes that support its prediction.

Interactive recognition is another promising use case. Rather than forcing a single prediction, a model can request a second view, a close-up of a proposed part, or confirmation of an attribute. The value of such interaction should be measured by uncertainty reduction per unit of user effort. This setting also changes the role of the backbone: a compact visual encoder with a calibrated uncertainty head may be preferable to a larger classifier that cannot indicate what additional evidence it needs.

8. Future research directions

A promising direction is evidence aggregation across modalities and views. A model could combine an image with a concise attribute description, a geolocation prior, or a second view while explicitly tracking which cue changes the prediction. Another direction is structured pretraining in which masked regions are reconstructed together with part relations, encouraging representations that preserve geometry rather than texture alone. For low-data categories, episodic evaluation and active annotation could focus human effort on the images that reduce uncertainty most.

Progress should also be measured with task-centered protocols. Instead of a single accuracy number, future studies should report robustness to background replacement, performance under class imbalance, calibration, energy per image, and explanation quality assessed through controlled perturbations. Such reporting would reward models that learn transferable evidence and discourage gains that depend on a narrow benchmark artifact.

Benchmark design itself also deserves attention. Existing datasets largely assume a fixed label set and a clean train-test division. More realistic suites could combine long-tailed sampling, novel-class discovery, temporal change, and multiple acquisition devices. They should retain a hidden evaluation split whose backgrounds and capture conditions differ from training, while documenting the semantic attributes that experts actually use. Such a benchmark would make it harder to obtain a high score through incidental correlations and would better reflect deployment conditions.

More principled compute allocation is another open direction. Rather than routing every image through an identical depth and token budget, a model can estimate the value of additional evidence. Easy examples with consistent global and local predictions may exit early, whereas ambiguous examples can request higher resolution or a second crop. The policy should be trained and evaluated jointly with accuracy so that speed gains do not merely shift errors to rare categories. Energy, memory traffic, and latency variance should be reported alongside throughput, especially for mobile and embedded deployment.

9. Conclusion

FGVC has evolved from supervised part localization to flexible systems that combine global context, selective local evidence, and increasingly capable pretraining. The architectural label matters less than the underlying allocation of information: a model must identify which regions matter, preserve relations among them, and remain cautious when the image does not support a confident distinction. Convolutional interaction modules, transformer token mechanisms, weak supervision, and parameter-efficient adaptation each address part of this challenge. The next stage of the field should be judged by cross-domain robustness, data efficiency, transparent evidence, and compute-aware behavior rather than by isolated gains on saturated benchmarks.

Disclosure statement

The authors declare no conflict of interest.

References

[1] Fu J, Zheng H, Mei T, 2017, Look Closer to See Better: Recurrent Attention Convolutional Neural Network for Fine-Grained Image Recognition. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 4438–4446.

[2] Yang Z, Luo T, Wang D, et al., 2018, Learning to Navigate for Fine-Grained Classification. Proceedings of the European Conference on Computer Vision (ECCV), 420–435.

[3] Luo W, Yang X, Mo X, et al., 2019, Cross-X Learning for Fine-Grained Visual Categorization. Proceedings of the IEEE/CVF International Conference on Computer Vision, 8242–8251.

[4] Dosovitskiy A, Beyer L, Kolesnikov A, et al., 2020, An Image Is Worth 16x16 Words: Transformers for Image Recognition at Scale. arXiv preprint arXiv:2010.11929.

[5] He J, Chen J, Liu S, et al., 2022, TransFG: A Transformer Architecture for Fine-Grained Recognition. Proceedings of the AAAI Conference on Artificial Intelligence, 36(1): 852–860.

[6] Xu Q, Wang J, Jiang B, et al., 2023, Fine-Grained Visual Classification via Internal Ensemble Learning Transformer. IEEE Transactions on Multimedia, 25: 9015–9028.

[7] Chen C, Fan Q, Panda R, 2021, CrossViT: Cross-Attention Multi-Scale Vision Transformer for Image Classification. Proceedings of the IEEE/CVF International Conference on Computer Vision, 357–366.

[8] Sun H, He X, Xu J, et al., 2024, SIM-OFE: Structure Information Mining and Object-Aware Feature Enhancement for Fine-Grained Visual Categorization. IEEE Transactions on Image Processing.

[9] Huang Y, Hechen Z, Zhou M, et al., 2025, An Attention-Locating Algorithm for Eliminating Background Effects

in Fine-Grained Visual Classification. IEEE Transactions on Circuits and Systems for Video Technology.

[10] Shi Y, Hong Q, Yan Y, et al., 2025, LDH-ViT: Fine-Grained Visual Classification Through Local Concealment and Feature Selection. Pattern Recognition, 161: 111224.

[11] Hinton G, Vinyals O, Dean J, 2015, Distilling the Knowledge in a Neural Network. arXiv preprint arXiv:1503.02531.

Publisher's note

Bio-Byword Scientific Publishing remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.