Journal on
Artificial Intelligence

Tech Science Press

https://doi.org/10.32604/jai.2026.085269

REVIEW

A Review of Fine-Grained Visual Categorization with Deep Learning

Richard Adusei[1,2,*] **and Gaddafi Abdul-Salaam**[2]

[1]Department of Computer Science, Faculty of Physical Sciences, University for Development Studies, Tamale, Northern Region, Ghana
[2]Department of Computer Science, College of Sciences, Kwame Nkrumah University of Science and Technology, Kumasi, Ashanti Region, Ghana

*Corresponding Author: Richard Adusei. Email: radusei@uds.edu.gh

Received: 08 May 2026; Accepted: 24 August 2026; Published: 14 September 2026

ABSTRACT: Fine-grained visual categorization (FGVC) presents a class of recognition problems in which the discriminative signal is spatially concentrated, visually subtle, and easily destroyed by the preprocessing and augmentation strategies that serve coarse recognition well. Where standard image classification requires a model to distinguish birds from cars, FGVC requires it to distinguish one bird species from another, a task that demands localization, feature-space shaping, and representation learning to operate in close coordination. This survey synthesizes thirty recent works spanning 2021 to 2026, organizing them under five interlocking themes, namely discriminative region discovery, metric learning and loss function design, data augmentation and training strategy, architectural and multi-scale feature representation, and foundation model adaptation. Rather than presenting these themes as independent research tracks, the survey examines them as overlapping responses to a single underlying difficulty, the impossibility of reliably separating where to look from how to represent what is found. Key tensions are identified across the literature between single-image and cross-image localization strategies, between background suppression and foreground discovery, between image-level and feature-level augmentation including generative diffusion augmentation, between geometrically principled and engineering-scale contrastive losses, and between the scale offered by foundation models and the architectural conservatism of purpose-built FGVC designs. The synthesis reveals that the most capable recent systems including FG-CLIP, FEM, HI2R, SIM-OFE, and CSDNet resolve these tensions through explicit integration of localization and feature-space objectives, that ultra-FGVC at the cultivar level is emerging as a distinct and harder sub-problem requiring dedicated solutions, and that the field's principal open problems including occlusion robustness, cross-dataset generalization, and the theoretical relationship between fine-grained discriminability and low-rank representation remain largely unaddressed.

KEYWORDS: Fine-grained visual categorization; vision transformer; CLIP; class-center loss; ultra-FGVC; multi-scale features; attention; data augmentation; deep learning

1 Introduction

Fine-grained visual categorization occupies an unusual position in computer vision. It is simultaneously one of the most practically important problems, distinguishing bird species, aircraft variants, and car models, and one whose core difficulty has resisted straightforward resolution by scaling data or model capacity alone. The central challenge is not that discriminative information is absent from images of, say, a Bushtit vs. a White-Breasted Nuthatch, but that it is spatially concentrated, visually subtle, and easily destroyed by the very preprocessing and augmentation techniques that help in coarser recognition tasks. A model that excels

Copyright © 2026 The Authors. Published by Tech Science Press. This work is licensed under a Creative Commons Attribution 4.0 International License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

at ImageNet classification may confidently fail on CUB-200-2011 because it has learned to recognize *bird* rather than *this particular bird*. The papers surveyed here collectively wrestle with that gap.

What emerges from reading these works together is not a single progressive narrative but rather a set of recurring tensions that different research groups resolve in different ways. Should discriminative regions be located by attending within a single image, or by comparing across images from the same category? Should training data be enriched at the image level or the feature level? Should large foundation models like CLIP be adapted to fine-grained tasks, or do their global pre-training objectives make them structurally ill-suited? These are not merely technical choices; they reflect deeper assumptions about where the difficulty of FGVC actually lives. The five themes identified in this synthesis are best understood not as separate research tracks but as overlapping responses to the same underlying question, *how do you force a model to notice the right thing?*

One particularly striking feature of this literature is how many papers span multiple themes. Ref. [1] is simultaneously an attention-based localization method and a contrastive learning proposal. Ref. [2] integrates saliency selection, graph-based relation modelling, and a novel loss function in a single framework. This cross-theme entanglement is not mere complexity for its own sake; it reflects a genuine insight, shared if rarely articulated, that region discovery and feature-space shaping are not separable problems. You cannot find the right regions without a discriminative feature space, and you cannot learn a discriminative feature space without attending to the right regions. The themes, then, are analytical rather than modular, and the most interesting intellectual action happens at their intersections.

2 Related Work

Several surveys have examined fine-grained visual recognition, and positioning the present synthesis among them is necessary for clarity and honesty about scope. Ref. [3] provided a broad taxonomy of FGVC methods covering part-based localization, attention mechanisms, metric learning, and data augmentation, but their coverage predates the transformer era and does not engage with the foundation model adaptation literature that defines much of the field in 2025 and 2026. Ref. [4] focused specifically on transformer-based part localization and offer a thorough treatment of that sub-problem, but treat loss function design and training strategy as peripheral rather than constitutive concerns. Ref. [5] surveyed both the problem setting and benchmark landscape comprehensively, with useful analysis of data collection methodology, but the architectural horizon is again CNN-dominated and the integration of vision-language models is not considered.

The present survey differs from these predecessors in three respects. First, it spans the 2021–2026 period specifically, covering the transition from CNN to Vision Transformer backbones and the emergence of CLIP-based adaptation as a distinct methodological category. Second, rather than organizing methods by mechanism, it organizes them by the conceptual tension they resolve, revealing cross-cutting relationships, that TransFG [1], ECC [6], and FEM [2] are jointly pursuing a discriminative feature space through different means and that this shared goal is obscured by a mechanism-based taxonomy. Third, the survey treats the relationship between region discovery and feature-space shaping as analytically central rather than as two independent problems, and uses this framing to explain why the most effective recent systems are also the most architecturally complex.

A note on scope and sources is warranted. The methods included here were selected as representative of the conceptual tensions the survey organises around, not as an exhaustive census of the field; the aim is analytical coverage of the recurring design problems rather than completeness over every published system. Because the 2025–2026 frontier moves quickly, several cited works were, at the time of writing, unrefereed preprints. These are flagged as “preprint” at first mention, and their single-paper results are reported as such

rather than as settled findings; synthesis-level claims are drawn from patterns across multiple works, not from any individual unrefereed result. Where a work has since appeared in a peer-reviewed venue, that version is cited.

3 Preliminaries

3.1 Problem Formulation

Fine-grained visual categorization is formally a multi-class classification problem defined over a set of visually subordinate categories. Given an input image x and a label space $Y = \{1, \ldots, C\}$ where all C classes belong to the same semantic superordinate (e.g., all bird species, all aircraft models, all car makes), the task is to learn a classifier $f: X \to Y$ that correctly identifies the subordinate category. The defining challenge is statistical, the between-class visual distance in subordinate recognition is small relative to the within-class visual distance. A Bushtit and a White-Breasted Nuthatch share most visual properties and differ only in specific, spatially localized attributes.

Formally, the class-conditional distributions $p(x|y = i)$ and $p(x|y = j)$ have high overlap for most pairs $(i, j) \in Y \times Y$, and the discriminative information is concentrated in a low-measure subset of the image domain whose location is not given at test time and varies across instances of the same class. This separates FGVC from object detection (where bounding boxes are the primary output) and standard transfer learning (where the difficulty is distributional shift rather than intrinsic within-target ambiguity). The literature surveyed here is exclusively concerned with image-level classification on established FGVC benchmarks, without relying on part annotations at training or test time.

3.2 Standard Backbones

The papers in this survey use three backbone families. Convolutional neural networks, principally ResNet-50 and ResNet-101, dominated through 2022 and remain relevant for augmentation-focused methods. Vision Transformers, principally ViT-B/16 pretrained on ImageNet-21K, became the dominant backbone from 2021 onward following TransFG and are used by the majority of post-2022 methods. The Swin Transformer, a hierarchical ViT variant using shifted-window attention, provides multi-scale feature maps that are particularly relevant for spatial and frequency domain methods. Foundation models, specifically CLIP trained on 400 million to 1.6 billion image-text pairs, introduce a qualitatively different backbone regime in which the pretrained model already encodes vision-language correspondences and the adaptation problem is fundamentally different from supervised fine-tuning of a vision-only encoder.

3.3 Evaluation Benchmarks

Performance in this literature is reported primarily on three benchmarks. CUB-200-2011 [7] contains 11,788 images across 200 North American bird species (5994 training/5794 test). Stanford Cars [8] contains 16,185 images across 196 car classes defined by make, model, and year. FGVC-Aircraft [9] contains 10,000 images of 100 aircraft model variants. Together, these three benchmarks cover biologically defined categories (birds), commercially defined categories (cars), and geometrically defined categories (aircraft), and methods that perform consistently across all three can be considered robustly general. Table 1 summarizes the six benchmark datasets most commonly used across the surveyed literature.

Table 1: Benchmark datasets used in fine-grained visual categorization research.

Dataset	Domain	Classes	Train	Test	Discrimination Challenge
CUB-200-2011	Birds	200	5994	5794	Plumage colour, wing markings, beak shape
Stanford Cars	Vehicles	196	8144	8041	Model year, trim level, badge detail
FGVC-Aircraft	Aircraft	100	6667	3333	Wing geometry, engine placement, fuselage length
NABirds	Birds	555	23,929	24,633	Similar to CUB-200 but larger and more ambiguous
Stanford Dogs	Dogs	120	12,000	8580	Coat texture, face structure, body proportion
iNaturalist	Multi-species	5089	~500K	~95K	Extreme class imbalance; real-world noise

4 Discriminative Region Discovery

Fine-grained visual categorization is, at its core, a localization problem before it is a classification problem. The discriminative signal that separates a Baird's Sparrow from a Clay-colored Sparrow, or a Boeing 747-100 from a 747-200, is spatially concentrated in a small, variable, and annotation-free subset of the image. The most direct response to this challenge is to build systems that explicitly locate discriminative parts without requiring expensive part annotations. Standard backbones, whether convolutional or transformer-based, process all spatial positions and treat them as equally relevant, which dilutes the discriminative signal into noise. The papers grouped under this theme share a common commitment to force the model to identify *where* to look without bounding-box or part-annotation supervision, pursuing that goal through three quite different strategies that reveal genuine conceptual disagreement about what it means for a region to be discriminative.

4.1 Single-Image Attention Localization

The most direct approach to discriminative region discovery uses the model's own internal attention or activation signals, computed over a single image, to identify informative spatial locations. The shared principle is that the model already contains the information needed to localize discriminative content; the task is to extract and exploit that information in a training-time or architecture-level loop.

Ref. [1] proposed TransFG, a ViT-based architecture that addresses the equal-treatment problem in standard patch tokenization. The method targets the identification of the K most discriminative patch token positions from the model's accumulated attention weights, using those positions to augment the global classification representation. TransFG computes an attention rollout matrix $A_{\text{final}} = \prod_{l=0}^{L-1} A_l$ by multiplying all per-layer attention matrices across the twelve transformer layers, yielding an accumulated importance score per patch. The top $K = 12$ highest-scoring token positions are selected, and their last-layer representations are concatenated with the `[CLS]` token to form the classification input $z_{\text{local}} = [z_{L-1}^0; z_{L-1}^{A_1}; \ldots; z_{L-1}^{A_K}]$. Overlapping patch tokenization (stride 12 vs. patch size 16) is also used to preserve inter-patch boundary information. While theoretically principled, multiplying all twelve layer attention matrices can propagate and amplify noise from early layers, making the final aggregated map less faithful to semantically discriminative

content than methods that draw on single-layer class attention signals. Fig. 1 illustrates the kind of inter-class ambiguity that TransFG is designed to resolve.

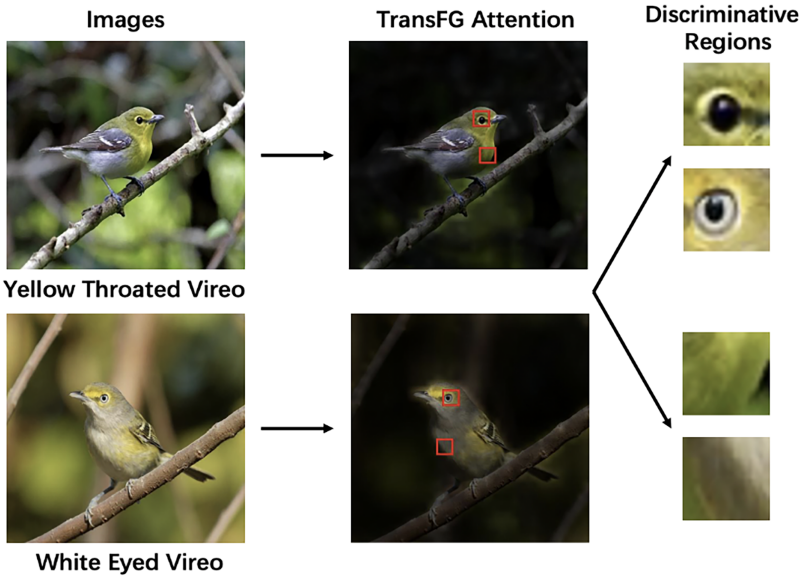

Figure 1: A confusing pair of instances from the CUB-200-2011 dataset, illustrating the fine-grained inter-class ambiguity that TransFG's attention-based token selection is designed to resolve [1].

Ref. [10] proposed FGVC-CAM, a lightweight alternative that uses Class Activation Maps (CAM) from a standard ResNet-50. The approach targets weak CAM localization as an anchor for data augmentation, improving discriminative coverage without modifying the backbone architecture. Spatial and channel attention are combined to derive a CAM over the input image. Three region-targeted augmentation operations then follow. Region Crop enlarges the identified discriminative region, Region Drop erases it entirely to expose secondary cues, and Region Mix blends the discriminative region of one image with the background of another. A joint loss is computed over both original and augmented samples. The quality of the CAM directly bounds the quality of the localization, creating a ceiling that ViT-based methods using richer attention signals do not share, and no ablation study is included to attribute gains to specific components.

Ref. [2] proposed the Feature Enhancement Module (FEM), a plug-and-play component that combines a Weakly Supervised Selector (WSS) and a dynamic Graph Convolutional Network (GCN). The aim is to select the most discriminative multi-scale features via a confidence-ranked selector and model structural relationships among selected parts using a dynamic GCN, enabling backbone-agnostic feature enhancement. A Feature Pyramid Network (FPN) extracts multi-scale feature maps which are fused into a unified representation; the WSS then ranks all spatial positions by a learned confidence score and retains the top-K most discriminative patches. A dynamic GCN Combiner constructs an input-adaptive adjacency matrix via scaled dot-product attention and applies graph convolution to model relational structure among selected patches. Top-M super-nodes are pooled through an MLP and global average pooling to yield the final classification descriptor. The top-K and top-M hyperparameters require dataset-specific tuning, and the GCN's adjacency matrix construction adds computational overhead beyond standard ViT inference.

Ref. [11] proposed SRL-FGVC, a cyclical attention framework designed to address the most under-explored failure mode in FGVC, that is heavy occlusion. The framework aims to force discriminative attention to spread across multiple spatial locations, including those a single forward pass would overlook, by

alternating mining and suppression cycles within a ViT backbone. SRL-FGVC interleaves two ViT-internal modules: the Mining Informative Attention Features (MIAF) module, which selects the top-K patches by aggregated attention weight and amplifies their features; and the Suppressing Attention using Margin Gradient (SAMG) module, which suppresses those tokens in the subsequent cycle with strength proportional to the classification gradient margin. A cooperative game mechanism adds an auxiliary branch trained on synthesized occluded inputs; a consistency loss $L_{\text{consist}} = \text{KL}(p_{\text{main}} \| p_{\text{aux}})$ enforces main-branch robustness to the occlusion patterns seen by the auxiliary branch. SRL-FGVC is the first FGVC method to directly target heavy occlusion as a training objective, but the cyclical mechanism's convergence and optimal iteration count are not analytically characterized, and the synthetic block-masking may not generalize to real-world occlusion where the occluding object has coherent texture and shape.

Ref. [12] proposed MARS-Trans (Multistage Attention Region Supplement Transformer), which augments a standard ViT-B/16 with three plug-in modules targeting three distinct failure modes simultaneously. The Multistage Attention Module (MAM) grades all H attention heads by their deviation from the mean head output and re-weights them to emphasize the heads that capture shallow intra-patch discriminative features: $\text{out}_{\text{MAM}} = \text{concat}(\beta_1 C_{all}[W_{max}],\ \beta_2 C_{all}[W_{min}],\ C'_{all})$. The Region Supplement Module (RSM), embedded in the final transformer layer, computes the cosine similarity of each patch token to the `[CLS]` token and supplements high-discriminability patches with their spatial neighbours' features to recover boundary information lost by non-overlapping tokenization, while blending low-discriminability patches with `[CLS]`. The Approximate Adjust Method (AAM) corrects classification logits at inference when the top-predicted class and the second-ranked class co-occur within the same training batch, a signal of approximate class confusion, the top-class logit is increased proportionally to its lead, resolving the confusion without any architectural modification. All three modules require no new parameters beyond one learnable scalar per layer. MARS-Trans achieves 91.9% on CUB-200-2011 and 95.3% on Stanford Cars with standard ViT-B/16, with the three-module combination contributing an average +1.8% over the baseline across benchmarks.

Ref. [13] proposed MPSA (Multi-Granularity Part Sampling Attention), which replaces anchor-box-based part localization with a fully-differentiable part-sampling mechanism operating at each of the four stages of a Swin-B backbone. Each Part Sampling Attention (PSA) module computes a learned softmax sampling map augmented by a learnable spatial bias B_i encoding dataset-specific spatial priors on where discriminative parts tend to appear, and uses it to sample a compact set of part feature vectors. A Multi-Granularity Part Retrospect (MPR) block applies PSA at every backbone stage and feeds the stage-specific part features upward to supplement the final high-level representation. A Grad-CAM supervision loss $L_{\text{cam}} = \|M_{\text{CAM}} - M^*\|_2$ aligns the learned feature weights with gradient-based discriminative maps, preventing the sampling process from collapsing to statistically convenient but non-discriminative regions. MPSA achieves 92.8% on CUB-200-2011, 95.4% on Stanford Cars, 95.4% on Stanford Dogs, and 92.5% on NABirds with Swin-B, with MPR adding +1.0% over single-scale PSA.

Ref. [14] proposed the Saccader framework (preprint), a biologically-inspired two-stage approach that replaces standard single-pass attention-based region discovery with a peripheral-to-fixation pipeline, reusing Wang et al.'s MPSA part-sampling mechanism as its feature encoder and building a novel priority map and saccadic fixation-sampling pipeline on top of it. The system first encodes a downsampled peripheral view of the entire image and generates a priority map S via the MPSA mechanism; then samples $N = 4$ non-redundant fixation patches from S using non-maximum suppression (Gaussian penalty kernel) to prevent spatial collapse, a systematic failure mode of part-based methods where selected patches cluster rather than covering distinct discriminative regions. High-resolution patches at each fixation point are extracted from the full-resolution source image and re-encoded by the same weight-shared backbone. A learned global impact factor $\alpha \in [0,1]$ weights fixation influence relative to the peripheral encoding, and per-fixation weights

β_n (low-temperature Boltzmann) modulate individual patch contributions via the fusion rule in Eq. (1):

$$z = \sigma\left(z_{per} + \alpha \cdot \frac{1}{N} \sum_n \beta_n z^n_{fix}\right) \quad (1)$$

A confidence-integrated negative log-likelihood loss supervises both the peripheral and fixation pathways jointly. The NMS non-redundancy constraint is the key structural novelty, specifically it guarantees that no two fixation patches are near-duplicates, providing a harder and more diverse set of discriminative patches than standard top-K selection. The Saccader achieves 91.8% on CUB with Swin-B and 97.7% on the EU-Moths insect dataset, with performance degrading significantly below $N = 2$ or above $N = 8$ fixations.

4.2 Relational and Graph-Structured Localization

Single-image attention methods identify important individual patch tokens but treat those patches independently. In fine-grained recognition, the discriminative signal often arises from the co-occurrence and structural relationship among multiple parts simultaneously. Relational and graph-structured methods address this by modelling higher-order dependencies among selected tokens.

Ref. [15] proposed HI2R (Hypergraph-guided Intra- and Inter-category Relation Modelling), which simultaneously addresses large intra-category variation and small inter-category differences. HI2R targets the modelling of high-order structural co-occurrence relations among discriminative patch groups within a category, and cross-category semantic relations via a persistent prototype cache, producing richer relational features than pairwise graph methods. After computing attention rollout over ViT-Base, HI2R selects the top-s patches by aggregated score and applies Fuzzy C-Means clustering to group them into $E = 6$ hyperedges with soft membership. A two-stage hypergraph convolution propagates structural information across the hyperedge set and fuses the result with the `[CLS]` token. The Inter-category Relation Perception (IRP) module maintains a per-class concept cache $P = \{p_c\}$ updated by moving average ($\alpha = 0.01$) across training batches, and applies cross-attention between the token sequence and this cache. A Random Semantic Consistency (RSC) loss enforces agreement between the attention map of the original image and its horizontal flip (see Eq. (7) below). The fixed $E = 6$ hyperedge count may not adapt to datasets with variable structural complexity, and the concept cache initialization is sensitive and adds memory overhead proportional to the number of classes.

Ref. [16] proposed SIM-OFE (Structure Information Mining and Object-Aware Feature Enhancement), which extends relational graph localization from co-occurrence modelling (HI2R) to geometric configuration modelling. The key insight is that the spatial arrangement among discriminative parts encodes additional species-level discriminative information beyond individual part activations. After a Hybrid Perception Attention (HPA) module fuses multi-head self-attention with a depth-wise convolutional branch via a learned gate to capture both long-range context and local texture, the Structure Information Mining (SIM) module selects the top-K discriminative part tokens by attention weight and constructs their polar coordinate representations (r_k, θ_k) relative to the image centroid. A graph $G = (V, E)$ is built with edge weights $w_{ij} = \exp(-\|\text{polar}_i - \text{polar}_j\|^2/\sigma^2)$ and a two-layer GCN propagates structural context. An Object-Aware Feature Enhancement (OFE) module fuses FPN multi-scale features with the GCN-refined part representations. The polar coordinate graph outperforms a plain Euclidean distance graph by +0.4% on CUB, confirming that the angular separation between parts carries information not captured by inter-part Euclidean proximity. SIM-OFE achieves 92.3% on CUB-200-2011, 96.1% on Stanford Cars, and 95.8% on FGVC-Aircraft. The centroid-based polar reference is not robust to off-centre cropping, where the image centroid does not coincide with the object centroid.

4.3 Cross-Image Co-Attention

A single image provides a single perspective on discriminative content, and single-image attention tends to fixate on the most prominent part. Cross-image co-attention methods address this by pairing same-category training images, exploiting the observation that regions mutually active across multiple images of the same class are more likely to be genuinely category-diagnostic.

Ref. [17] proposed PCA-Net (Progressive Co-Attention Network), which pairs same-category training images and uses channel-wise bilinear co-attention to identify mutually shared discriminative channels. The network aims to overcome single-region fixation by mining all category-diagnostic channels via iterative cross-image comparison, building a diverse multi-part representation without part annotations. Given a same-class image pair (x_a, x_b), the Channel Interaction (CA) Module computes a bilinear similarity matrix over channel-wise descriptors to identify channels mutually active in both images; attention maps are re-weighted to emphasize these shared channels. An Attention Erasure (AE) Module then zeros out the attended regions before the next co-attention round, forcing successive rounds to discover complementary regions. Multiple progressive rounds of CA → AE → CA yield a diverse set of part attention maps whose feature vectors are aggregated for classification. A centre loss additionally penalizes intra-class scatter. At inference time, co-attention across a pair is not applicable; the model falls back to single-image attention with pre-trained weights, meaning the cross-image reasoning is entirely a training-time signal. Fig. 2 illustrates this iterative pipeline.

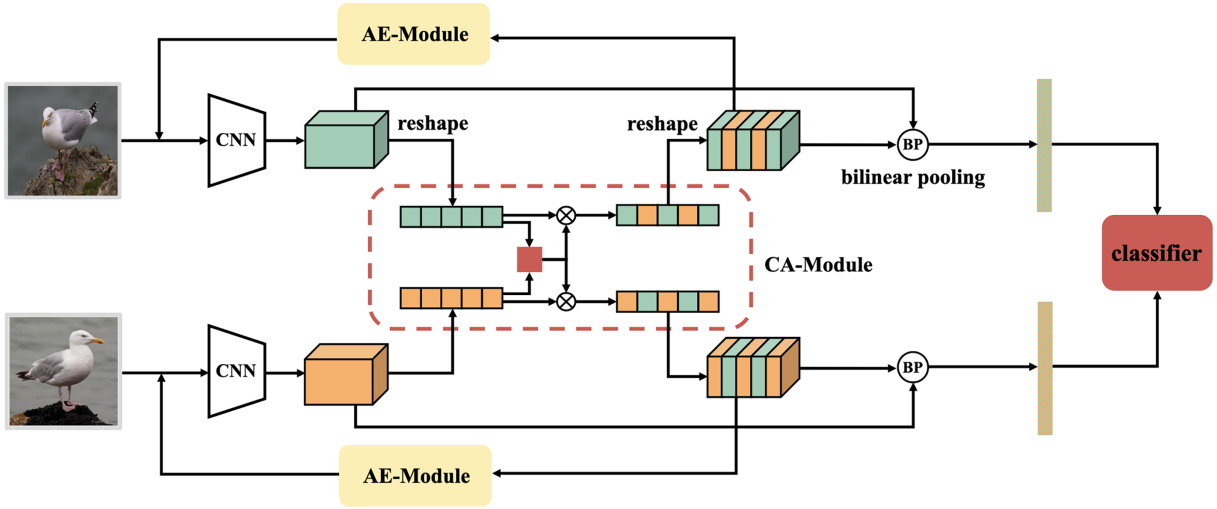

Figure 2: The framework of the progressive co-attention network (PCA-Net), showing the Channel Interaction (CA) and Attention Erasure (AE) modules used iteratively to discover complementary discriminative regions [17].

4.4 Background Suppression for Ultra-FGVC

Standard discriminative region discovery assumes the foreground subject dominates the image. In ultra-FGVC, where inter-class differences are sub-cultivar-level (e.g., different soybean cultivars, cotton varieties), background variation can dominate the discriminative signal, causing models to latch onto irrelevant context rather than the subtle foreground differences. Methods in this sub-theme directly model and suppress background content as part of the localization objective.

Ref. [18] proposed SV-Transformer, a Swin-B-based architecture with two complementary modules targeting ultra-FGVC's background problem. The Object Feature Extractor (OFE) inserts a global pooling

token x_g into the patch sequence and progressively selects the top-$k = 16$ most discriminative patches by their attention scores to x_g, fusing them with x_g via a cross-attention block to produce a focused object feature Z_f. The Background Perturbation Learning (BPL) module identifies background patches as those falling below the 20th percentile of the global-token attention score distribution, then synthesizes hard training images I' by replacing those background patches with patches sampled from a distractor pool. The model is trained jointly on original and perturbed images under a weighted sum of classification and background-perturbation losses, $L = \alpha \cdot L_{\text{cls}} + \beta \cdot L_{\text{bpl}}$. BPL encodes a structural prior absent from standard attention methods: a foreground-focused model should classify correctly even when the background is replaced by an arbitrary distractor. Ablation confirms BPL provides a larger improvement on ultra-FGVC datasets where background variation is high (Cotton80, SoyCultivar) than on standard FGVC (CUB). SV-Transformer achieves Cotton80 72.08%, SoyLocal 65.17%, SoyGene 88.83% and CUB-200-2011 91.3%.

4.5 Taxonomy

Fig. 3 organises the eleven surveyed methods into four sub-themes according to the scope of their localization signal: single-image attention localization (1a), relational and graph-structured localization (1b), cross-image co-attention (1c), and background suppression for ultra-FGVC (1d). The primary axis of variation is whether discriminativeness is defined within a single forward pass over foreground content, within a structured graph over patch groups, by comparison across images of the same class, or by explicit modelling and suppression of non-discriminative background.

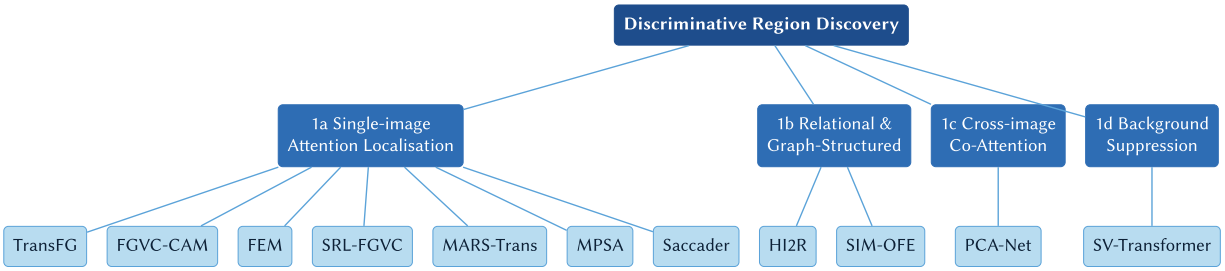

Figure 3: Taxonomy of discriminative region discovery methods (eleven surveyed papers across four sub-themes): TransFG [1], FGVC-CAM [10], FEM [2], SRL-FGVC [11], MARS-Trans [12], MPSA [13], Saccader [14], HI2R [15], SIM-OFE [16], PCA-Net [17], SV-Transformer [18].

4.6 Synthesis of Discriminative Region Discovery

The eleven papers in this theme collectively reveal a deepening understanding of what it means for a region to be "discriminative" in FGVC. The first group defines discriminativeness through single-image or cross-image activation signals. PCA-Net operationalises discriminativeness through shared channel activation across same-category images, defining a region as discriminative when it consistently triggers identical channels across multiple instances. TransFG replaces cross-image comparison with within-model temporal aggregation, treating a region as discriminative if cumulative attention across all twelve layers consistently highlights it. FGVC-CAM adopts a behavioural criterion by classifying a region as discriminative only when its removal measurably degrades classification performance. FEM introduces a confidence-based threshold, where a region qualifies as discriminative if it exceeds a saliency level following multi-scale fusion. SRL-FGVC relies on a resilience criterion, identifying regions as discriminative only if correct classification persists after occluding all previously detected areas.

A second group reframes discriminativeness as a relational, geometric, or structural property rather than a per-patch score. HI2R shifts to a relational perspective by treating discriminativeness as an emergent

property of patch *groups* rather than isolated patches. SIM-OFE incorporates a geometric dimension, demonstrating that the *angular arrangement* between discriminative parts carries species-level information independent of individual part features. MARS-Trans exploits attention-head heterogeneity by identifying the most deviant heads as encoders of shallower, spatially specific features. MPSA grounds discriminativeness in gradient signals, validating parts as discriminative only when Grad-CAM confirms their direct contribution to the classification decision. The Saccader introduces a non-redundancy constraint, enforcing spatial dispersion among discriminative patches by applying non-maximum suppression to the priority map as a direct implementation of the diversity criterion. Finally, the SV-Transformer inverts this paradigm entirely by shifting focus from foreground discriminativeness to background identification, explicitly training the model to become invariant to non-discriminative regions.

These eleven definitions produce complementary failure modes. Attention rollout (TransFG, HI2R) propagates early-layer noise. CAM-based localization (FGVC-CAM, MPSA) is bounded by a circular dependency between localization quality and classifier quality. Cross-image co-attention (PCA-Net) cannot be applied at inference. The polar-coordinate graph in SIM-OFE fails when the subject is off-centre. The Saccader's non-differentiable NMS sampler cannot be trained end-to-end. SV-Transformer's background quantile threshold is fixed and dataset-agnostic. Crucially, the theme now spans two regimes: standard FGVC (CUB, Cars, Aircraft) and ultra-FGVC (Cotton80, SoyCultivar), with SV-Transformer, CSDNet, and CLA-Net (the latter two in Themes 3 and 2, respectively) forming a nascent ultra-FGVC sub-literature whose methods differ architecturally and empirically from species-level FGVC. Table 2 consolidates all eleven methods, cross-referencing each paper's core problem, method, principal strength, principal weakness, and evaluation metric.

Table 2: Reference summary for Theme 1: Discriminative region discovery (eleven methods, four sub-themes).

Reference	Core Problem	Method	Strength	Weakness	Metric
TransFG [1]	Uniform patch-token treatment	PSM: product of all layer attention matrices; append top-K token representations to `[CLS]`	First to aggregate all-layer ViT attention; overlapping patches preserve local boundaries	Rollout noise accumulates across 12 layers; strongly reliant on ImageNet-21K pretraining	Top-1 accuracy
FGVC-CAM [10]	Generic augmentation destroys discriminative cues	Region Crop, Drop, Mix guided by spatial + channel attention CAM on ResNet-50	Lightweight; no architectural changes; forces secondary discriminative region discovery	CAM quality bounds localization; no ablation study	Top-1 accuracy
FEM [2]	Insufficient inter-class feature separability	FPN + WSS top-K selection + dynamic GCN Combiner + Class-Centric Loss	Plug-and-play with CNN and ViT; adaptive saliency ranking	Top-K, top-M require per-dataset tuning; GCN adds overhead	Top-1 accuracy
SRL-FGVC [11]	Attention collapse under occlusion	Cyclical MIAF + SAMG in ViT + cooperative occluded auxiliary branch	First FGVC method targeting heavy occlusion; discovers multiple regions	Synthetic block-masking may not transfer to real occlusion; cycle count uncharacterised	Top-1 accuracy

(Continued)

Table 2 (continued)

Reference	Core Problem	Method	Strength	Weakness	Metric
MARS-Trans [12]	Head uniformity dilutes shallow features	MAM head-deviation weighting + RSM patch boundary supplement + AAM logit correction; ViT-B/16	Three orthogonal improvements; no new parameters; +1.8% avg	MAM threshold α heuristic; AAM adds inference overhead	Top-1 accuracy
MPSA [13]	Anchor-box localization loses part shape	PSA softmax sampling + learnable bias B_i + MPR multi-stage retrospect + L_{cam} Grad-CAM loss; Swin-B	Differentiable; multi-granularity; no part annotations; CUB 92.8%	Four PSA stages memory-intensive; biases must be re-tuned per dataset	Top-1 accuracy
Saccader [14]	Spatial redundancy in part selection	Peripheral priority map (reused MPSA encoder) + NMS fixation sampling ($N = 4$) + weight-shared re-encoding + confidence NLL	NMS prevents patch clustering; weight-shared encoder; strong insect results	Non-differentiable sampler; fixed scale factors; non-FGVC baseline encoder	Top-1 accuracy
HI2R [15]	Pairwise graphs miss high-order co-occurrence	HSL: Fuzzy C-Means hyperedges + hypergraph conv; IRP: concept cache + cross-attention	Hyperedges capture multi-patch co-occurrences; IRP uses cross-category semantics	Fixed $E = 6$ hyperedges; cache initialization sensitive	Top-1 accuracy
SIM-OFE [16]	Part positions ignore inter-part geometry	HPA hybrid attention + SIM polar-coord GCN over top-K parts + OFE FPN fusion; ViT	Polar graph +0.4% over Euclidean; captures angular configuration between parts	Centroid reference fails on off-centre crops; GCN depth limited to 2	Top-1 accuracy
PCA-Net [17]	Single-image attention fixates on single region	CA Module (bilinear channel co-attention) + AE Module (progressive erasure) across same-class pairs	Discovers complementary regions via cross-image channel alignment	Cross-image reasoning is training-time only; CNN backbone	Top-1 accuracy
SV-Transformer [18]	Background dominates ultra-FGVC signal	OFE global-token top-$k = 16$ selection + BPL background-patch distractor swap; Swin-B	Explicitly suppresses background; hard samples at no annotation cost; ultra-FGVC SOTA	Hard top-k; distractor pool curation needed; Swin-B only	Top-1 accuracy

5 Metric Learning and Loss Function Design

Locating discriminative regions addresses the spatial problem in FGVC, but the complementary challenge is geometric: even when the right regions are found, the learned feature space may cluster different classes too closely and scatter instances of the same class too widely to support reliable classification. The papers in this theme therefore intervene not at the spatial level of localization but at the level of the feature space itself, introducing loss functions and learning objectives that explicitly shape the geometry of the representation. Contrastive learning methods operate on instance pairs and push specific representations apart or together. Class-centre methods maintain running estimates of per-class means and optimise directly for compactness around those means. Adversarial discrimination methods use a min-max game to enforce domain-level invariance while preserving class-level discriminability.

5.1 Contrastive Learning

Contrastive learning trains a model by simultaneously pulling representations of similar instances closer and pushing representations of dissimilar instances apart, without reference to a fixed class prototype. For FGVC, the critical question is which pairs to contrast.

The contrastive component of [1] introduced a pairwise contrastive feature loss alongside the Part Selection Module of TransFG. The contrastive component targets intra-class compactness and inter-class separability in the last-layer feature space via a margin-based loss applied to classification representations. The loss is computed over all same-class and different-class pairs in the batch (Eq. (2)):

$$L_{\text{con}} = \frac{1}{B^2} \sum_i \left[\sum_{j:\, y_i = y_j} \left(1 - \text{Sim}(z_i, z_j)\right) + \sum_{j:\, y_i \neq y_j} \max\left(\text{Sim}(z_i, z_j) - \alpha,\ 0\right) \right] \tag{2}$$

where $\text{Sim}(\cdot)$ is cosine similarity and $\alpha = 0.4$ is a margin. Both objectives are active simultaneously throughout training. The $O(B^2)$ pairwise computation is effective but the global margin α does not adapt to the relative difficulty of specific class pairs.

Ref. [19] proposed Con-Swin (Contrastive Swin-Transformer), which uses the Swin Transformer's hierarchical multi-scale feature extraction as the backbone and adds a lightweight two-layer projection head. The objective is to decouple the contrastive and classification objectives by routing the contrastive loss through a separate projection head, allowing the backbone to learn discriminative representations without the contrastive objective distorting backbone features. Swin-B processes the image through four hierarchical stages; outputs from all four stages are concatenated into a unified multi-scale feature $z = \sigma(\text{Concat}(z^1, z^2, z^3, z^4))$. A projection head $\tilde{z} = \text{LN}(\text{FC}_2(\sigma(\text{FC}_1(z))))$ produces abstract representations used for contrastive loss. A decisive ablation shows that applying the contrastive loss directly to backbone features gives only +0.07% improvement, while routing it through the projection head gives +1.48%, replicating the SimCLR insight that projection heads prevent the contrastive objective from distorting backbone feature space. The hyperparameter β must be tuned per dataset.

Ref. [20] proposed FG-CLIP, which extends contrastive learning to the regional level by constructing a 12M-image fine-grained hard-negative dataset (FineHARD). The training regime aims to scale CLIP's contrastive pre-training to fine-grained visual-linguistic correspondences by combining high-quality detailed recaptions, region-level supervision via RoIAlign, and hard negatives generated by modifying specific visual attributes. A two-stage pipeline first recaptions 1.6 billion image-text pairs with CogVLM2-19B to produce entity-rich long descriptions (up to 248 tokens), training with global InfoNCE. Stage 2 uses FineHARD with 40M bounding boxes and 10M attribute-modified hard negative captions, training with $L = L_{\text{global}} + 0.1 \cdot L_{\text{regional}} + 0.5 \cdot L_{\text{hard}}$, where L_{regional} uses RoIAlign to extract K region embeddings matched to phrase-level

text embeddings. Stage 1 training requires approximately 30 days on 160 NPUs, placing it beyond the reach of most research groups.

5.2 Class-Center Push/Pull

Class-center methods maintain running estimates of per-class feature means and construct loss functions that pull each instance toward its own class center while pushing it away from the centers of competing classes. This is a fundamentally different geometric objective from contrastive learning: instead of shaping pairwise distances, it shapes the relationship between instances and the distributional structure of their class.

Ref. [6] proposed ECC (Exploration of Class Center), a plug-in loss combining a Multiple Class-Center Constraint (MCC) and a Class-Center Label Generation (CLG) module. The goal is to improve feature space geometry by pulling instances toward their class center and pushing them from the most similar non-target class center, while replacing one-hot labels with soft labels derived from class-center similarity distributions to mitigate overfitting. Class-center feature vectors F_y are maintained via online moving-average updates. The MCC loss is given by Eq. (3):

$$L_{\text{MCC}} = \sum_{k} \left(1 - \cos(X_k, F_{y_k}) + s_{y_k, \text{sim}_{y_k}} \cdot \cos(X_k, F_{\text{sim}_{y_k}})\right) \tag{3}$$

The CLG module converts class-center distributions into soft probability labels Q_{y_i} and computes a KL divergence loss. The final loss is $L_{\text{final}} = L_{\text{CE}} + \lambda_1 L_{\text{MCC}} + \lambda_2 L_{\text{CLG}}$. No architectural changes are required. The push component can be unstable early in training when class-center estimates are zero-initialised and unreliable. Fig. 4 provides an overview of the ECC framework.

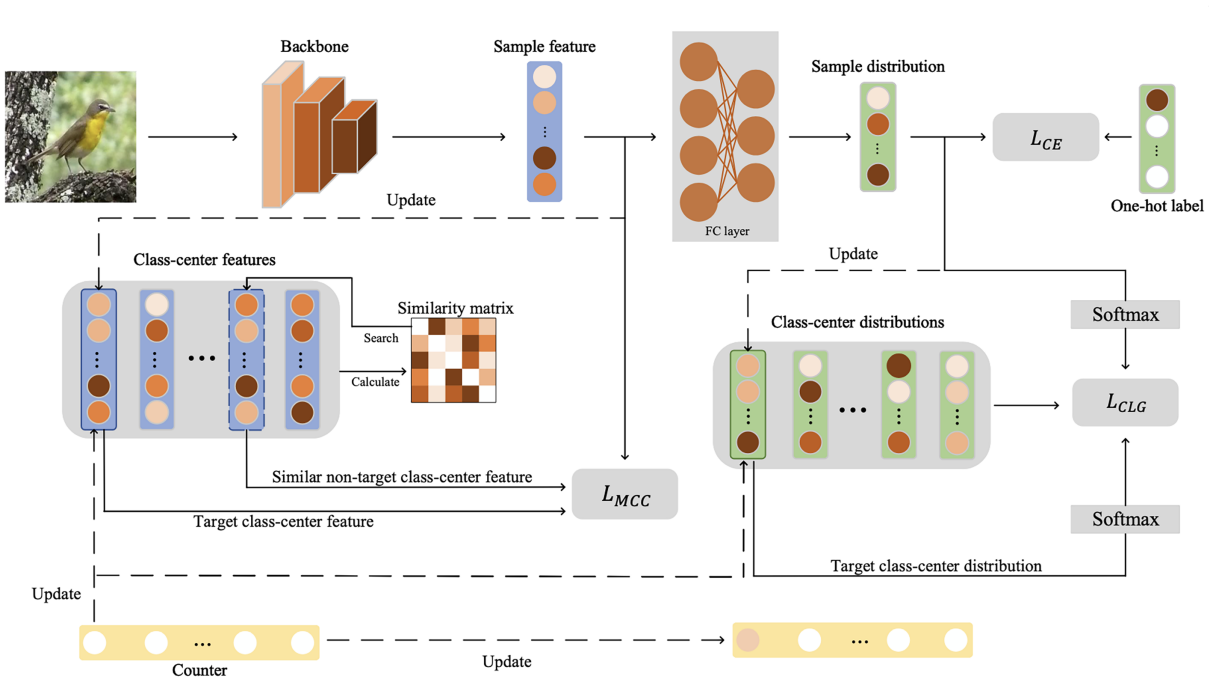

Figure 4: Overview of ECC, illustrating the Multiple Class-Centre Constraint (MCC) and Class-Centre Label Generation (CLG) modules that jointly shape the feature space geometry [6].

Ref. [2] independently arrived at a class-center-based loss in FEM's Class-Centric Loss (CCL), using Pearson correlation as the similarity metric. The class-centric loss targets feature space geometry around

dynamically maintained class centers using a metric that captures linear feature co-variation rather than treating feature dimensions independently. CCL pulls features toward their target class center using Pearson correlation (L_{feature}), and soft labels are derived from class-center distributions via Jensen-Shannon divergence for a complementary self-distillation loss (L_{class}). Ablation confirms that Pearson correlation marginally outperforms cosine similarity on most datasets, and that the dynamic GCN only marginally outperforms static GCN (+0.17%), suggesting the CCL training signal dominates the GCN's structural modelling contribution.

Ref. [21] proposed DTRG (Dynamic Target Relation Graph), which extends the class-centre principle from individual centre push/pull to structural supervision over the *entire inter-class similarity graph simultaneously*. DTRG maintains class centres via exponential moving-average updates, constructs a dynamic adjacency matrix $A_{ij} = \cos(C_i, C_j)$ over all class centres at every training step, and applies a Graph Similarity Loss (Eq. (4)) that minimises the Frobenius norm between the learned feature-space similarity matrix and a self-supervised target relation matrix T derived from label co-occurrence within each batch:

$$L_{\text{GSL}} = \|A - T\|_F^2 \tag{4}$$

An Online Centre Loss (OCL) additionally minimises $\sum_i \|f_i - C_{y_i}\|^2$ for intra-class compactness. The combined loss $L = L_{\text{CE}} + \lambda_1 L_{\text{OCL}} + \lambda_2 L_{\text{GSL}}$ requires no architectural changes and plugs in as a regulariser. The GSL's key advantage over MCC (ECC) and CCL (FEM) is that it shapes all C^2 inter-class relationships simultaneously rather than only the target-vs.-most-confusable-class pair. DTRG achieves CUB-200-2011 90.1%, Stanford Cars 95.7%, and FGVC-Aircraft 94.1% with ResNet-101, with GSL contributing +0.8% in isolation and the dynamic centre update outperforming static centres by +0.6%.

The centre loss in [17] provides a fourth, simpler incarnation of the class-center principle. It enforces intra-class compactness in the joint representation produced by the progressive co-attention mechanism: $L_{\text{center}} = \sum_i \|x_i - c_{y_i}\|_2^2$, penalising the Euclidean distance between each sample's feature and its class centre. Centers are updated online per batch. This pull-only formulation enforces intra-class compactness without explicitly pushing away from competing centres, which is precisely the missing inter-class component that ECC's MCC and FEM's CCL were designed to address.

Ref. [22] proposed CLE-ViT (Contrastive Learning Encoded Vision Transformer), which targets ultra-FGVC's central challenge: standard augmented positives are too similar to drive useful contrastive gradients when intra-class variation already rivals inter-class variation. The core contribution is a self-supervised module that constructs *hard positive pairs* from a single image by simultaneously shuffling spatial blocks (2×2 grid) and masking a variable fraction $\alpha \in [0.15, 0.45]$ of patches, producing a structurally disrupted view that shares class identity with the original. A triplet loss with margin $\beta = 1$ then enforces large inter-class distance between the anchor, this hard positive, and a randomly sampled negative from the same batch. The final loss is $L = L_{\text{clsa}} + \lambda L_{\text{clsp}} + \gamma L_{\text{icl}}$ with $\lambda = \gamma = 1$ (tuned to $\lambda = 0.3$, $\gamma = 0.5$ for standard FGVC). CLE-ViT achieves a +19% improvement over the second-best method on SoyGlobal, the most dramatic single-paper delta in the surveyed ultra-FGVC literature, while contributing only a modest +0.6% on CUB-200-2011. The magnitude gap directly validates the core hypothesis: the harder the intra-class variation, the more valuable hard positive pairs become.

Ref. [23] proposed CLA-Net (Contrastive Lie Algebra Network), which performs contrastive learning in *Lie algebra space* using RGB covariance matrices as second-order feature representations, directly addressing the overfitting risk of instance-level contrastive methods in the extreme low-data regime of ultra-FGVC (3–11 training images per class). For each image, the 3×3 RGB covariance matrix $V = \frac{1}{S-1} \sum_k (x_k - u)(x_k - u)^T$ forms a Symmetric Positive Definite (SPD) matrix lying on the SPD manifold, a Riemannian symmetric space $GL^+(d)/O(d)$ rather than a Lie group. The matrix logarithm $T = \log_{I_d}(V)$ maps V to the tangent

space at the identity, the Lie algebra of $GL(d)$, where standard Euclidean distance is geometrically valid on the curved SPD manifold. A lightweight FC module predicts T from the image features. The contrastive loss (Eq. (5)) combines MSE between two views' Lie algebra representations and a prediction error term:

$$L_{ss} = \sum_{i=1}^{P} \left((T_i^A - T_i^{B'})^2 + |T_i - T_i'| \right) \tag{5}$$

with $\lambda = 0.03$. The Lie algebra module is active only during training and discarded at inference. CLA-Net achieves Cotton80 62.08%, SoyLocal 52.17%, and Stanford Cars 94.1% (ResNet-50, 2K features). The SPD manifold constraint is the key principled advance: because Lie algebra contrastive constraints are softer than instance-level congruence, the method discourages overfitting while still maximising inter-class separation.

5.3 Adversarial Discrimination

Ref. [24] proposed WDAN (Weighted Discriminative Adversarial Network), which extends open-set domain adaptation to fine-grained scenarios. The system aims to transfer fine-grained classification knowledge to a target domain with shared and unknown classes, preventing unknown-class samples from contaminating the shared-class alignment during adversarial training. A Selective Transformer Encoder (STE) uses overlapping patch tokenisation to preserve local fine-grained details. Two independent classification heads provide complementary discriminative signals; their disagreement on a target sample serves as a proxy for unknown-class membership. A Weighted Conditional Domain Discriminator (WCDD) assigns sample weights $w_i = 1 - P(\text{unknown}|x_i)$, effectively excluding likely-unknown samples from alignment gradients. WDAN additionally introduces the WebStanfordCars benchmark for fine-grained open-set domain adaptation. The approach's main vulnerability is that weight estimation depends on dual-classifier calibration; if both classifiers agree on a wrong prediction for an unknown-class sample, that sample will be incorrectly aligned.

5.4 Taxonomy

Fig. 5 organises the ten surveyed methods into three sub-themes according to the geometric relationship they impose in feature space: contrastive learning (instance-level repulsion/attraction), class-centre push/pull (cluster-level and graph-level geometry), and adversarial discrimination (domain-level alignment). The primary axis of variation is the granularity at which the loss operates: instance pairs or manifold-level (2a), class centroids or inter-class graphs (2b), or domain distributions (2c).

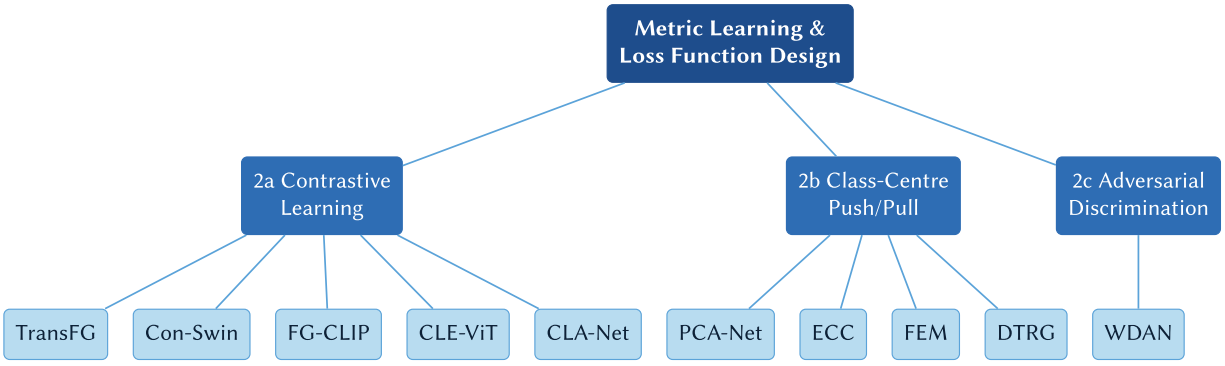

Figure 5: Taxonomy of metric learning & loss function design methods (ten surveyed papers across three sub-themes): TransFG [1], Con-Swin [19], FG-CLIP [20], CLE-ViT [22], CLA-Net [23], PCA-Net [17], ECC [6], FEM [2], DTRG [21], WDAN [24].

5.5 Synthesis of Metric Learning and Loss Function Design

The most striking structural pattern in this theme is the convergence of *four* independent research groups on class-centre-based supervision, each arriving via different technical routes. PCA-Net uses a pull-only Euclidean centre loss; ECC adds an MCC inter-class push targeting the most confusable class; FEM re-derives an equivalent push-pull via Pearson correlation; and DTRG escalates this from individual class pairs to the entire inter-class similarity graph, minimising the Frobenius norm between the learned and target-relation matrices simultaneously over all C^2 pairs. The fact that ECC, FEM, and DTRG converge on related structures from different starting points strongly suggests that class-centre geometry is a generic and robust training signal for FGVC.

The contrastive learning sub-theme now spans three distinct regimes of difficulty and data availability. TransFG and Con-Swin target standard FGVC with whole-image features; FG-CLIP targets region-level correspondences with billion-scale data. CLE-ViT and CLA-Net target ultra-FGVC with 3–11 training images per class, a data regime where standard augmented positives are too easy and instance-level feature congruence constraints overfit catastrophically.

Within the ultra-FGVC regime, the two methods diverge in how they harden the contrastive signal. CLE-ViT's shuffle+mask hard positive construction directly addresses this challenge by operating on the principle that positives must be structurally disrupted to yield meaningful gradients in low-data settings. CLA-Net's Lie algebra contrastive operates one level of abstraction higher, replacing feature-level congruence with covariance-manifold proximity, and the SPD manifold geometry ensures the constraint is softer, maximising inter-class separation without enforcing exact view congruence. The +19% ultra-FGVC gain (CLE-ViT on SoyGlobal) and the Lie algebra ablation advantage over instance contrastive (CLA-Net on SoyLocal) collectively establish that the ultra-FGVC contrastive problem is qualitatively different from the standard FGVC contrastive problem and requires dedicated solutions.

Con-Swin's ablation remains one of the most informative experiments in the surveyed literature. It demonstrates that routing features through a projection head yields a +1.48% gain over direct backbone contrastive learning (+0.07%), confirming that SimCLR's projection-head insight generalises to FGVC. Specifically, separating the contrastive and classification objectives proves more critical than the exact formulation of the contrastive loss. Table 3 summarises the methods surveyed in this theme alongside their core problem, method, strength, weakness, and evaluation metric.

Table 3: Reference summary for metric learning & loss function design.

Reference	Core Problem	Method	Strength	Weakness	Metric
TransFG [1]	Weak subcategory inter-class separability	Cosine-similarity margin loss over all same/different-class pairs in batch	Directly applied to confusing subcategory pairs; no extra data or architecture	$O(B^2)$ cost; global margin α does not adapt to pair difficulty	Top-1 accuracy
Con-Swin [19]	Contrastive loss distorts backbone features	Swin-B four-stage concat + 2-layer projection head; L_{CE} on backbone + L_{con} on head	Projection head prevents distortion; +1.48% vs. +0.07% without head	β requires per-dataset tuning	Top-1 accuracy

(Continued)

Table 3 (continued)

Reference	Core Problem	Method	Strength	Weakness	Metric
FG-CLIP [20]	Global CLIP alignment neglects regional attributes	1.6B LMM-recaptioned pairs + FineHARD RoI regional contrastive + LLM hard-negative loss	Regional alignment + hard negatives scale to fine-grained discrimination	~30 days on 160 NPUs; FineHARD proprietary	Top-1 accuracy
CLE-ViT [22]	Standard augmented positives too easy for ultra-FGVC	Swin-B + self-supervised shuffle + mask hard positive + triplet loss ($\beta = 1$)	Hard positives from single image; +19% on SoyGlobal; no extra annotation	Random batch negatives; small gain on standard FGVC (+0.6% CUB)	Top-1 accuracy
CLA-Net [23]	Instance contrastive overfits in 3–11 sample regime	RGB covariance SPD → matrix-log Lie algebra; MSE contrastive + prediction loss; $\lambda = 0.03$	Geometrically valid Euclidean metric on SPD manifold; soft constraint prevents overfitting	Only $d = 3$ RGB covariance; no hard negative mining; ResNet-50 only	Top-1 accuracy
ECC [6]	Inter-class confusability geometry ignored	MCC (cosine push-pull on most-confusable pair) + CLG (KL soft labels from centre distributions)	Plug-in; backbone-agnostic; simultaneously compresses intra- and separates inter-class	Zero-initialised centres may destabilise early; λ_1, λ_2 tuning	Top-1 accuracy
FEM [2]	CE lacks geometric inter-class separation	CCL: Pearson correlation push-pull + JS-divergence soft labels	Pearson captures linear co-variation; JS soft labels more principled than uniform smoothing	λ_1, λ_2, λ_3 tuning; risk of overfitting on small datasets	Top-1 accuracy
DTRG [21]	Pairwise losses ignore global inter-class structure	Dynamic cosine graph over EMA centres + $L_{GSL} = \|A - T\|_F^2$ + OCL; no architecture changes	Shapes all C^2 pairs simultaneously; self-supervised target T; complementary to attention methods	Batch-derived T noisy at small batch size; CNN-only evaluation	Top-1 accuracy
PCA-Net [17]	Insufficient intra-class feature compactness	Standard Euclidean centre loss on joint co-attention features	Simple; no extra data; +0.4%–0.6% over CE only	No inter-class push component; clusters can shrink without separating	Top-1 accuracy
WDAN [24]	Unknown classes corrupt domain alignment	STE (overlapping-patch ViT) + dual classifiers + WCDD weighted adversarial alignment	First FG-OSDA benchmark (WebStanfordCars); prevents negative transfer	Dual-classifier calibration fragile; limited to visually close domain pairs	H-score

6 Data Augmentation and Training Strategy

Standard augmentation techniques are designed to enforce invariances useful for coarse recognition, but can erase exactly the fine-grained cues that FGVC relies on. Standard data augmentation for image classification enforces spatial, scale, and photometric invariances that are genuinely useful for coarse-grained recognition, yet for FGVC these same techniques are systematically harmful: random cropping removes discriminative parts; random colour jitter erases plumage colour differences; Mixup blends two images including their backgrounds, creating training examples where the discriminative features of two different subcategories coexist at arbitrary mixing ratios in a single image, producing label-inconsistent composites. The papers in this theme share a common commitment: augment in a way that increases training diversity without destroying the discriminative cues that FGVC relies on.

6.1 Image and Region-Level Augmentation

The most direct response to the specificity problem is to anchor augmentation operations to identified discriminative regions rather than the whole image.

Ref. [10] addressed the problem that standard augmentation randomly destroys discriminative regions by anchoring the Crop, Drop, and Mix operations to regions identified by a class activation map. The method aims to improve localisation and feature extraction by designing region-targeted augmentation that enforces discovery of multiple discriminative cues. CAM-derived attention localises the discriminative region per training image; Region Crop enlarges the identified region; Region Drop erases it entirely, forcing the model to find alternative discriminative cues; Region Mix blends the discriminative region of one image with the background of a different-class image, training the model to classify based on discriminative content rather than background context. A combined loss $L = L_{\text{ori}} + L_{\text{crop}} + L_{\text{drop}} + L_{\text{mix}}$ sums contributions from original and all augmented views. In early training, the CAM will be noisy and the augmented regions will be misidentified, introducing corrupted training signal.

Ref. [25] proposed MRDMN (Multiresolution Discriminative Mixup Network), which addresses the specific failure mode of Mixup in FGVC known as manifold intrusion. The network targets this failure mode by restricting blending to discriminative regions and deriving area-proportional soft labels, alongside a multi-resolution knowledge distillation pipeline. A Soft-CAM localisation pipeline computes and smooths class activation maps, then applies morphological opening (erosion followed by dilation) to isolate compact discriminative blobs. The Discriminative Mixup Module (DMM) blends only the discriminative region pixels between image pairs; soft labels are derived proportionally to the discriminative area contributed by each source image rather than using the fixed λ, making the label semantically consistent with spatial content. Multiple resolution branches (448×448, 224×224, 112×112) are trained simultaneously, and a Multiresolution Distillation step transfers knowledge from all teacher branches to a single-resolution student network via KL-divergence losses. Soft-CAM quality degrades early in training, and the three-stage pipeline substantially increases training cost.

6.2 Feature-Level Semantic Augmentation

Image-level augmentation, however carefully region-targeted, modifies pixels and can accidentally destroy fine-grained discriminative cues. Feature-level augmentation operates after the backbone has encoded discriminative content, augmenting in the feature space along directions that correspond to semantic variation.

Ref. [26] proposed LSDA (Learnable Semantic Data Augmentation), a meta-learning framework that trains a Covariance Prediction Network (CovNet) to predict sample-wise feature covariance matrices as

semantic directions for implicit feature-level augmentation. The framework aims to augment deep feature representations along sample-specific semantic directions in feature space, bypassing the discriminative region destruction problem of pixel-level operations and preventing CovNet degenerate collapse via meta-learning. For each training sample x_i with deep feature a_i, CovNet predicts a diagonal covariance matrix $\Sigma_i^g = g(a_i; \theta_g)$ representing the local semantic variation around that sample. The ISDA loss (Eq. (6)) upper-bounds the expected cross-entropy over translations along the $\mathcal{N}(0, \lambda\Sigma_i^g)$ distribution:

$$\ell^{\mathrm{ISDA}}(x_i, y_i) = -\log \frac{e^{w_{y_i}^T a_i + b_{y_i}}}{\sum_j e^{w_j^T a_i + b_j + \frac{1}{2} v_{jy_i}^T \Sigma_i^g v_{jy_i}}} \tag{6}$$

A meta-learning algorithm trains CovNet and the classifier alternately, ensuring CovNet produces covariances that genuinely improve generalisation rather than trivially minimising the ISDA loss by outputting near-zero matrices. Meta-learning adds approximately 3× the standard training cost per iteration, and CovNet predicts only the diagonal of the covariance matrix, discarding potentially important off-diagonal co-variation structure.

6.3 Consistency Regularisation

Consistency regularisation augments training by requiring the model's output to be invariant under transformations that should not affect the classification decision.

Ref. [15] introduced the Random Semantic Consistency (RSC) loss in HI2R as a structural regulariser on attention maps. The loss targets attention-map overfitting to single prominent regions by regularising attention against flip-consistent views. For each training image x, HI2R computes attention maps M on both the original and the horizontally flipped version; the RSC loss (Eq. (7)) minimises:

$$L_{\mathrm{RSC}} = \frac{1}{HW} \sum_{i,j} \left\| M_{i,j} - \mathrm{Flip}(M')_{i,j} \right\|_2 \tag{7}$$

This enforces bilaterally symmetric consistency in the attention map, preventing the model from focusing on a single prominent but non-bilateral region. The RSC loss requires only an additional forward pass on the flipped image with no extra parameters, yet contributes +0.56% accuracy improvement in isolation. The flip-consistency assumption may be unreliable for categories with strongly asymmetric discriminative features.

6.4 Curriculum and Hardness-Aware Training

Curriculum methods modify the order and weighting of training data based on the model's assessed difficulty with different category pairs.

Ref. [27] proposed PS-GHD (Progressive Self-Guided Hardness Distillation), a three-stage curriculum distillation pipeline. The curriculum aims to prevent teacher model error propagation by separating hard and easy categories and applying stage-adaptive training signals that prevent hard-class confusion from interfering with easy-class learning. Stage 1 (Ancestor) trains a ViT-Base end-to-end on the target dataset using standard cross-entropy; hard categories are identified from the Ancestor's confusion matrix. Stage 2 (Parent) trains with temperature-scaled reverse KL distillation ($T = 10$) from the Ancestor, but relabels hard categories under a single shared "hard class" label, preventing the Parent from being distracted by indistinguishable hard-class boundaries. Stage 3 (Child) applies stage-adaptive loss switching: easy categories receive standard label-smoothed cross-entropy; hard categories receive distillation from the Parent's soft

predictions via reverse KL. The three-stage pipeline requires approximately 3× training time relative to a single-stage method, and the binary hardness criterion is a coarse simplification of what is genuinely a continuous and instance-dependent difficulty distribution.

Ref. [28] proposed CSDNet (Contrastive Self-Distillation Network), which attacks ultra-FGVC's limited-data challenge at three levels simultaneously. At the instance level, the Subcategory-Specific Discriminative Parsing (SSDP) module identifies the largest connected discriminative region via a binarised class activation map and generates a cropped foreground sample I' that spotlights subcategory-specific content. At the feature level, a Dynamic Discrepancy Learning (DDL) module maintains a FIFO memory queue of length ι and applies a contrastive loss over both current and historical features: positive pairs are pulled together, negative pairs are pushed apart up to margin ξ. At the logit level, the Subcategory-Specific Discrepancy Transfer (SSDT) module trains two branches (raw and augmented) jointly and distils augmented-branch knowledge into the raw branch via $L_{\text{ssdt}} = \text{KL}(y_{\text{raw}} \| y_{\text{aug}})$; at inference, only the raw branch is used, achieving 113.67 FPS without sacrificing the discriminative knowledge learned from augmented views. CSDNet achieves Cotton80 67.92%, SoyLocal 60.50%, and SoyGene 86.86% with Swin-B. The SSDT's inference-time elimination of the augmented branch is the sharpest practical contribution: it brings augmentation-based training benefits at zero inference cost, a principle applicable beyond ultra-FGVC.

6.5 Generative Data Augmentation

Generative models offer a qualitatively new augmentation paradigm: instead of transforming existing training images, they synthesise entirely new images conditioned on class and structural priors, providing diversity that deterministic augmentations cannot.

Ref. [29] proposed SaSPA (Structure and Subject Preserving Augmentation), a five-step diffusion-based pipeline that addresses the fidelity-diversity trade-off of existing generative augmentation methods. Text-to-image methods generate diverse but often class-incorrect images; Img2Img methods use real images as guidance and produce near-duplicates with limited diversity. SaSPA resolves this by *decoupling* structural conditioning from appearance conditioning: a Canny edge map extracted from the source image provides structural constraints via ControlNet, while a BLIP-diffusion subject reference drawn from a *different* same-class image provides appearance conditioning, the source image itself is never used directly. LLM-generated text prompts (100 per dataset) diversify context, lighting, and orientation. Generated images are filtered by CLIP semantic similarity and by predictive confidence (top-$k = 10$ predictions of a pre-trained baseline classifier). An augmentation ratio α (probability of replacing a real sample with a synthetic one per epoch) controls the proportion of synthetic data; empirically, optimal α scales inversely with real data size, making synthetic data most valuable in ultra-FGVC's extreme low-data regime. SaSPA achieves 93.8% on Stanford Cars, 86.6% on FGVC-Aircraft, and 83.2% on CUB-200-2011 (ResNet-50), outperforming all traditional and generative baselines across five datasets.

6.6 Taxonomy

Fig. 6 organises the surveyed methods into five sub-themes according to the level at which the augmentation or training intervention operates: image/region-level manipulation, feature-level semantic augmentation, consistency regularisation over attention maps, curriculum or hardness-aware scheduling, and generative augmentation.

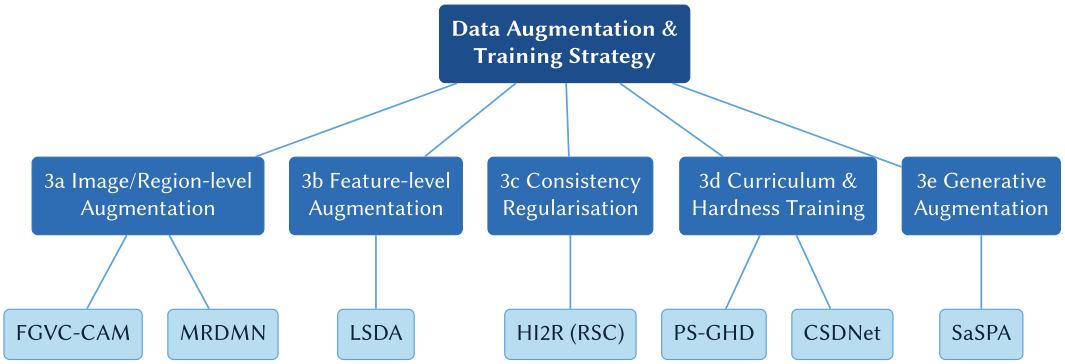

Figure 6: Taxonomy of data augmentation & training strategy methods (eight surveyed papers across five sub-themes): FGVC-CAM [10], MRDMN [25], LSDA [26], HI2R (RSC) [15], PS-GHD [27], CSDNet [28], SaSPA [29].

6.7 Synthesis of Data Augmentation and Training Strategy

The papers in this theme collectively map out the landscape of the specificity problem: the tension between augmentation's goal of increasing diversity and FGVC's requirement of preserving discriminative cues. They reveal that there are at least five distinct levels at which augmentation or training structure can intervene: the pixel level (FGVC-CAM, MRDMN), the feature level (LSDA), the attention-structure level (HI2R's RSC loss), the training curriculum level (PS-GHD, CSDNet), and the synthesis level (SaSPA). The progression from pixel-level to feature-level to generative augmentation reflects a clear conceptual trajectory: pixel-level operations, however carefully region-targeted, cannot guarantee that fine-grained cues are preserved; feature-level augmentation sidesteps this by operating after the backbone encodes discriminative structure; generative synthesis goes further still, producing entirely new images that can diversify both appearance and background.

MRDMN's area-proportional soft label is the sharpest technical contribution in the image-level sub-theme and generalises well beyond FGVC: the standard Mixup label λ assigns credit proportionally to the mixing coefficient, which is not semantically grounded; MRDMN replaces this with a label proportional to the discriminative area contributed by each source, which is semantically grounded. The consistency regularisation sub-theme is anchored by HI2R's RSC loss, which enforces bilateral symmetry of attention maps across horizontal flips at almost no parameter cost, delivering +0.56% in isolation. This suggests that attention-map consistency is a systematically underexploited signal in transformer-based FGVC.

CSDNet introduces a novel dimension largely absent from the other papers: ultra-FGVC at inference efficiency. By combining spatial-scale decoupled pre-training, dynamic difficulty learning, and self-supervised distillation transfer, CSDNet achieves 113.67 FPS on Cotton80, a practically deployable system, not only an accuracy-focused research prototype. SaSPA occupies a different niche: unlike all other methods in this theme that transform or re-weight existing training images, SaSPA synthesises new images from diffusion priors, raising an important question about the boundary between data augmentation and dataset construction. The observation that optimal mixing weight α scales inversely with real data size suggests that generative augmentation is most valuable precisely where FGVC is hardest: few-shot and ultra-FGVC regimes with 3–11 images per class. Table 4 lists the methods surveyed in this theme together with their core problem, method, strength, weakness, and evaluation metric.

Table 4: Reference summary for data augmentation & training strategy.

Reference	Core Problem	Method	Strength	Weakness	Metric
FGVC-CAM [10]	Generic augmentation destroys discriminative cues	CAM-guided Crop, Drop, Mix; joint loss on original + augmented samples	Lightweight; forces secondary discriminative region discovery	CAM quality unreliable early in training; no ablation study	Top-1 accuracy
MRDMN [25]	Region-unaware Mixup corrupts sample labels	Soft-CAM + morphological opening then DMM region Mixup; area-proportional soft labels; multi-resolution distillation	Avoids manifold intrusion; area-proportional labels semantically consistent; five-dataset evaluation	Complex three-stage pipeline; ResNet only; CAM noise in early training	Top-1 accuracy
LSDA [26]	Pixel augmentation corrupts discriminative regions	CovNet (MLP predicting diagonal covariance) + ISDA implicit feature augmentation + meta-learning	Feature-level augmentation preserves discriminative regions; architecture-agnostic; consistent +1%–2% gains	3× training cost; diagonal-only covariance; off-diagonal co-variation ignored	Top-1 accuracy
HI2R (RSC) [15]	Attention overfits to single region	RSC loss: $\|M - \text{Flip}(M')\|_2$; no extra parameters	No parameters; cheap forward pass; +0.56% in isolation	Assumes bilaterally symmetric discriminative features	Top-1 accuracy
PS-GHD [27]	Hard-class confusion propagates to students	Ancestor (CE) then Parent (RKL, hard class relabelling, $T = 10$) then Child (adaptive LS/RKL switching)	Prevents teacher error propagation; differentiated treatment improves both easy and hard categories	3× training time; binary hardness criterion ignores continuous difficulty	Top-1 accuracy
CSDNet [28]	Ultra-FGVC: overfitting + slow inference on cultivar-level data	SSDP (spatial-scale decoupled pre-training) + DDL (dynamic difficulty learning) + SSDT (self-supervised distillation transfer at logit level)	State-of-the-art on Cotton80 (67.92%); 113.67 FPS inference; three-level curriculum targets ultra-FGVC specifically	Method components add training complexity; evaluated only on ultra-FGVC datasets	Top-1 accuracy

(Continued)

Table 4 (continued)

Reference	Core Problem	Method	Strength	Weakness	Metric
SaSPA [29]	Insufficient intra-class diversity for rare fine-grained categories	ControlNet (edge-map conditioning) + BLIP-diffusion (subject-reference generation); optimal α scales inversely with real data size	Synthesises entirely new images rather than transforming existing ones; 93.8% on Stanford Cars	Proprietary diffusion backbone; synthesis quality depends on subject-reference fidelity; α requires per-dataset tuning	Top-1 accuracy

7 Architecture and Multi-Scale Feature Representation

The first three themes intervene within a fixed backbone; this theme addresses the backbone itself, asking what architectural choices determine whether a model can capture the multi-scale, multi-domain, and structurally complex representations that FGVC requires. The architectural choices in this literature reflect a field in transition from CNN to transformer backbones, and from single-scale to multi-scale representation. Standard single-scale architectures process all image content at a uniform spatial resolution, which is suboptimal for FGVC where discriminative features span multiple scales simultaneously: global shape context at coarse scale, part configuration at medium scale, surface texture at fine scale.

7.1 Multi-Domain Spatial and Frequency Fusion

CNN feature maps operate in the spatial domain, capturing coarse shape and texture through convolutional filters. Fine-grained discriminative cues such as feather barbule periodicity or surface microtexture are naturally represented as mid-to-high frequency components in the Discrete Cosine Transform domain, which spatial convolutions capture only implicitly and incompletely.

Ref. [30] proposed SFFF (Spatial-Frequency Feature Fusion), a dual-branch CNN architecture that processes the same image in both the spatial RGB domain and the DCT frequency domain. The architecture targets the capture of complementary frequency-domain discriminative cues and learns to selectively exchange informative channels between the two branches. The first branch processes the RGB image through a standard ResNet; the second branch converts the image to DCT coefficients (8×8 block DCT, YCrCb representation). An importance-sparsity selection mechanism ranks all frequency feature channels by a learned attention score and retains the top-K informative channels. The RSEF module then ranks channels by discriminative importance, selects the top-K, exchanges uninformative channels between branches (identified via batch-normalisation scaling factors), and fuses the result via 1×1 convolution. A global-local branch structure provides both full-image and CAM-cropped patch features, ensembled at inference. The dual-branch structure doubles inference FLOPs, and DCT applied at a fixed 8×8 block level is a non-learnable transform. The finding that frequency features benefit FGVC-Aircraft and Stanford Cars more than CUB-200-2011 suggests that frequency utility correlates with the regularity and periodicity of discriminative textures in the target category.

7.2 Efficient Attention Design

Standard ViT self-attention scales as $O(N^2 D)$ per layer, which becomes prohibitive for FGVC's high-resolution inputs (448×448 producing $N \approx 784$ tokens). Efficient attention methods replace the full $O(N^2)$

interaction with structured approximations that reduce complexity while preserving the model's ability to capture discriminative long-range dependencies.

Ref. [31] proposed DDA-Trans (Dual-Dependency Attention Transformer), which decomposes standard self-attention into two parallel linear-complexity pathways. The aim is to reduce attention complexity from $O(N^2)$ to $O(mN)$ while separately capturing positional-structural and semantic-content dependencies, with a discriminative ability enhancement module refining the `[CLS]` token with the most discriminative patches at each layer. A Center Token Aggregation (CTA) module projects and aggregates the N patch tokens into $m = 16$ representative centre tokens. Position Dependency Attention (PDA) captures local positional structure by dividing tokens into $g = 48$ groups with local windowed attention and global cross-window attention via centre tokens. Semantic Dependency Attention (SDA) applies bidirectional cross-attention between patch tokens and centre tokens. Both pathways are fused with learnable scalar weights. A Discriminative Ability Enhancement (DAE) module computes `[CLS]`-to-patch attention scores and uses the top-$t = 8$ discriminative patches to refine the `[CLS]` token representation. Total complexity reduces from $O(N^2D)$ to $O(2gND + 6mND + 4ND^2)$, linear in N for the attention terms. The non-standard choice of 11 transformer layers instead of the canonical 12 gives DDA-Trans a FLOPs advantage not attributable to the dual-path mechanism, undermining the efficiency comparison.

7.3 Hierarchical Multi-Scale Aggregation

Single-scale ViT representations process all patches at a uniform spatial resolution. Hierarchical multi-scale aggregation methods extract features at multiple spatial resolutions and combine them to produce representations sensitive to fine-grained structure at all relevant scales.

Ref. [19] used Con-Swin's multi-scale aggregation as a deliberate architectural choice: exploiting the Swin Transformer's existing four-stage hierarchical structure and concatenating all four stage outputs without any additional cross-scale fusion modules. The multi-scale component targets Swin's built-in hierarchy, providing multi-scale context via simple stage concatenation without additional fusion modules: $z = \sigma(\text{Concat}(z^1, z^2, z^3, z^4))$. This concatenated feature z serves as the backbone representation for both the classification head and the projection head used for contrastive loss. The approach requires no additional modules, fusion losses, or stage-specific supervision. Direct concatenation without explicit cross-scale alignment may not fully exploit complementary scale-specific information, but the strong results obtained support the observation that backbone choice often matters more than method-level innovations.

Ref. [2] provided a more structured approach to multi-scale aggregation in FEM's Feature Pyramid Network (FPN). The FPN targets a unified, discriminative multi-scale representation by fusing spatial information from all backbone stages with learned, input-adaptive weights. Rather than concatenating all scales, the FPN computes a weighted fusion of features from all backbone stages into a single unified representation, with fusion weights learned to balance scale contributions. Ablation confirms that FPN alone contributes +0.80% over the baseline, making it the single largest contributor among FEM's components.

7.4 Knowledge Distillation Across Scales

Distillation-based multi-scale methods extract information at multiple resolutions or training stages but deliver it to a single compact network at inference time, avoiding test-time multi-scale processing costs.

Ref. [25] in MRDMN trained separate 448×448, 224×224, and 112×112 resolution branches during training and distilled their collective knowledge into a single student network. The distillation component transfers fine-to-coarse complementary multi-resolution knowledge to a single-scale inference network, avoiding the test-time cost of multi-branch processing. Each resolution branch shares a common ResNet-101

backbone trunk with an independent classification head; during training, KL-divergence distillation losses transfer knowledge from all teacher branches to the student. At inference, only the full-resolution branch is used.

Ref. [27] in PS-GHD transferred not scale-complementary but hardness-differentiated knowledge across training stages. The pipeline transfers hardness-differentiated supervision from earlier training stages, preventing teacher confusion from propagating and providing stage-appropriate training signals for hard and easy categories. Reverse KL divergence (mode-seeking) is used for distillation in both Stage 2 and Stage 3 rather than forward KL (mean-seeking), because mode-seeking KL produces sharper distributions better suited for fine-grained discrimination. Temperature $T = 10$ softens the teacher distribution sufficiently to provide informative gradients for hard-category boundaries.

7.5 Hierarchical Label-Guided Multi-Granularity

Scientific taxonomy provides a natural label hierarchy that organises categories from coarse to fine (order → family → genus → species). Most FGVC methods ignore all but the finest (species) level, discarding supervision signal that is freely available for bird, aircraft, and plant benchmarks.

Ref. [32] proposed CHBC (Cross-Hierarchical Bidirectional Consistency) (preprint), which exploits the full taxonomic tree by learning orthogonal multi-granularity features and enforcing bidirectional probability consistency across hierarchy levels. The Multi-Granularity Enhancement (MGE) module generates one attention map A_i and feature map F_i per hierarchy level i. For level $i \geq 2$, it removes the projection of level-i features onto the coarser level-$(i-1)$ representation via an orthogonalisation operation (Eq. (8)) analogous to Gram-Schmidt projection:

$$\mathrm{MOD}(M_{\mathrm{fine}}, M_{\mathrm{coarse}}) = M_{\mathrm{fine}} - \frac{\sum M_{\mathrm{fine}} \odot M_{\mathrm{coarse}}}{\sum M_{\mathrm{coarse}} \odot M_{\mathrm{coarse}}} M_{\mathrm{coarse}} \tag{8}$$

The enhanced feature $\hat{F}_i = F_i' \odot A_i'$ retains only the incremental discriminative information beyond the coarser granularity. The Cross-Hierarchical Bidirectional Consistency (CBC) loss uses the transition matrix $D_{i,j} = \prod_{k=i}^{j-1} D_{k,k+1}$ to project fine-level classification scores upward (coarse prediction) and coarse-level scores downward (fine prediction), then minimises Jensen-Shannon divergence between projected and direct predictions at each level: $L_{\mathrm{con}} = \sum_{l=1}^{h} \mathrm{JS}(s_l, \hat{s}_l)$. Total loss is $L = L_{\mathrm{cls}} + L_{\mathrm{con}}$. CHBC achieves weighted accuracy of 90.4% on CUB-200-2011 (species 87.8%), 95.3% on FGVC-Aircraft, and 95.4% on Stanford Cars with a ResNet-50 backbone. Ablations confirm that both MGE and CBC contribute independently. The key limitation is the requirement for hierarchical label annotations beyond the target species level, which are available on standard FGVC benchmarks but absent from ultra-FGVC cultivar datasets such as Cotton80.

7.6 Taxonomy

Fig. 7 organises the eight surveyed methods into five sub-themes according to the architectural mechanism used to capture multi-scale or multi-granularity features.

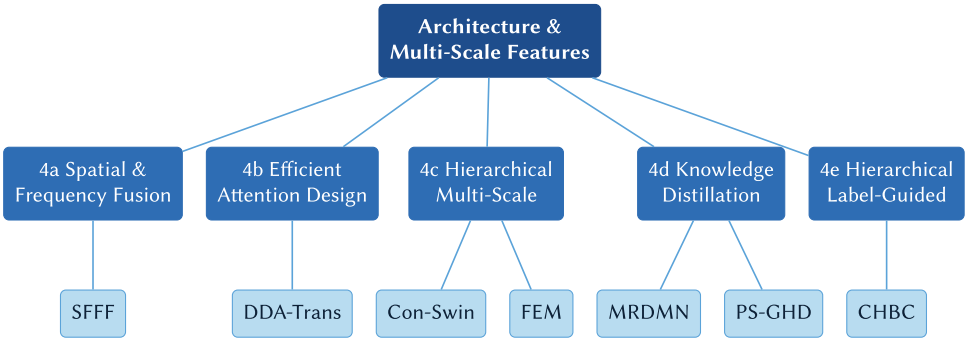

Figure 7: Taxonomy of architecture & multi-scale feature representation methods (eight surveyed papers across five sub-themes): SFFF [30], DDA-Trans [31], Con-Swin [19], FEM [2], MRDMN [25], PS-GHD [27], CHBC [32].

7.7 Synthesis of Architecture and Multi-Scale Feature Representation

The architectural trajectory within this theme reflects the broader shift from CNN to transformer backbones and from single-scale to multi-scale representations. SFFF is a CNN-era method that identifies a genuine and architecturally motivated gap: spatial CNN features miss frequency-domain discriminative cues that are physically present in the image. The finding that frequency features benefit FGVC-Aircraft and Stanford Cars more than CUB-200-2011 suggests that frequency utility correlates with the regularity and periodicity of discriminative textures in the target category, providing actionable guidance for dataset-specific architecture selection. Yet SFFF evaluates only on ResNet backbones and does not explore whether the frequency-domain intuition applies to transformer architectures, where the patch tokenisation already implicitly encodes some frequency structure through learned position embeddings.

DDA-Trans proposes an efficiency modification to multi-head attention, addressing spatial token complexity ($O(N^2)$ scaling with input resolution) by introducing centre token aggregation to reduce the effective sequence length. Its efficiency argument is undermined by a non-standard 11-layer architecture, which gives it a FLOPs advantage unrelated to the dual-path mechanism itself. The contrast between Con-Swin's minimal four-stage concatenation and FEM's structured FPN + GCN pipeline illustrates a recurring tension in the FGVC literature between elegant simplicity and structured complexity. FEM's FPN contributes the largest single ablation gain (+0.80%), but the GCN adds only +0.76% with substantially more complexity, suggesting that multi-scale fusion is the key architectural decision rather than the relational modelling among selected features.

CHBC introduces a qualitatively different axis of multi-scale reasoning: rather than scaling across spatial resolutions, it scales across semantic abstraction levels defined by external taxonomic labels. The Gram-Schmidt-inspired orthogonalisation in MGE is a principled decomposition, each hierarchy level's features capture only the discriminative information not already encoded at the coarser level, directly addressing the feature entanglement that arises when fine-level classifiers implicitly re-learn coarser-level patterns. The bidirectional CBC loss ensures that the model's classification scores are mutually consistent across levels, eliminating the contradictory multi-level predictions that can arise from independent per-level classification heads. The key open question is applicability: the taxonomy hierarchy is readily available for standard FGVC benchmarks (CUB, Aircraft, Cars) but absent from ultra-FGVC cultivar datasets, limiting CHBC's reach. This highlights a structural limitation of taxonomic approaches: their value is proportional to the richness and depth of the available label hierarchy. Table 5 summarises the methods surveyed in this theme alongside their core problem, method, strength, weakness, and evaluation metric.

Table 5: Reference summary for architecture & multi-scale feature representation.

Reference	Core Problem	Method	Strength	Weakness	Metric
SFFF [30]	Spatial CNNs miss frequency textures	Dual CNN (RGB + DCT/YCrCb) + importance-sparsity selection + RSEF adaptive channel exchange	Frequency features complement spatial; forced branch collaboration via BN-based channel switching	Dual-branch doubles inference FLOPs; DCT is non-learnable; no ViT evaluation	Top-1 accuracy
DDA-Trans [31]	Quadratic attention conflates positional-semantic dependencies	CTA centre tokens + PDA (local + global) + SDA (patch ↔ centre) + DAE top-t `[CLS]` refinement	Linear-in-N complexity; explicit position/semantic decomposition	Non-standard 11-layer ViT gives unfair FLOPs advantage; still underperforms leading methods	Top-1 accuracy
Con-Swin [19]	Single-scale ViT lacks multi-granularity features	Swin-B four-stage output concatenation $z = \sigma(\text{Concat}(z^1, z^2, z^3, z^4))$	Minimal implementation; no extra modules; backbone multi-scale structurally motivated	No cross-scale alignment; result illustrates backbone choice dominates method-level innovation	Top-1 accuracy
FEM [2]	Single-scale maps miss multi-scale structure	FPN multi-scale fusion with learnable weights then WSS then dynamic GCN Combiner	Plug-and-play; FPN alone contributes +0.80%	Global fusion weights not input-adaptive; overhead from FPN + WSS + GCN combined	Top-1 accuracy
MRDMN [25]	Single-resolution training lacks scale complementarity	Three-resolution teacher branches (448/224/112) then KL distillation to single student	Complementary cross-resolution knowledge; five-dataset evaluation; single-scale inference	Complex three-stage training; CAM quality bounds distillation; ResNet only	Top-1 accuracy
PS-GHD [27]	Undifferentiated distillation spreads teacher errors	Ancestor then Parent (hard relabelling, RKL, $T = 10$) then Child (adaptive LS/RKL switching)	Prevents teacher error propagation; differentiated easy/hard supervision	3× training time; binary hardness criterion too coarse	Top-1 accuracy

(Continued)

Table 5 (continued)

Reference	Core Problem	Method	Strength	Weakness	Metric
CHBC [32]	Species-level training ignores coarser taxonomy	MGE Gram-Schmidt orthogonalisation per level + CBC JS-divergence bidirectional consistency ($L = L_{cls} + L_{con}$)	Exploits freely available taxonomy; orthogonal MGE features disentangle granularity levels; CUB 90.4%, Aircraft 95.3%	Requires hierarchical annotations absent in ultra-FGVC datasets; ResNet-50 backbone; weighted accuracy metric not directly comparable to standard top-1	Weighted accuracy

8 Foundation Model Adaptation

The most recent literature reflects a shift in the field's assumptions: if large pre-trained models like CLIP already encode rich visual and linguistic representations, the question becomes not how to build representations from scratch but how to specialise pre-trained representations for fine-grained discrimination. The papers in this theme therefore represent a qualitatively different assumption from the previous four themes. Rather than developing methods that train discriminative representations from supervised labels on established FGVC benchmarks, they assume that foundation models, specifically CLIP trained on 400 million to 1.6 billion image-text pairs, already encode rich visual and linguistic representations, and that the primary challenge is specialising those representations for fine-grained discrimination. This shifts the question from how to design a better attention mechanism or loss function to how to close the gap between CLIP's global image-text alignment objective and the regional, subtle, attribute-level discriminability required for fine-grained visual understanding.

8.1 CLIP Adaptation for Fine-Grained Vision-Language Alignment

CLIP's pre-training objective aligns whole-image representations with short, coarse-grained text descriptions. This global alignment is useful for zero-shot classification of broad categories but fails on fine-grained tasks because the discriminative visual signal is concentrated in local regions that coarse captions do not describe.

Ref. [20] proposed FG-CLIP, which addresses CLIP's coarse-grained alignment problem at scale. The training regime aims to scale CLIP's contrastive pre-training to fine-grained visual-linguistic correspondences by combining high-quality detailed recaptions, region-level supervision via RoIAlign, and semantically-grounded hard negatives generated by modifying specific visual attributes. A two-stage training pipeline is used. Stage 1 recaptions 1.6 billion image-text pairs with CogVLM2-19B to produce entity-rich long descriptions (up to 248 tokens), and trains with global InfoNCE on the CLIP ViT-L/14 architecture. Stage 2 constructs FineHARD: 12M source images from GRIT, with 40M bounding boxes detected by Yolo-World and 10M attribute-modified hard negative captions generated by Llama-3.1-70B (changing one attribute: colour, shape, texture). Stage 2 trains with $L = L_{global} + 0.1 \cdot L_{regional} + 0.5 \cdot L_{hard}$, where $L_{regional}$ uses RoIAlign to extract K region embeddings matched to phrase-level text embeddings. The enormous scale of Stage 1 (approximately 30 days on 160 NPUs) and the proprietary FineHARD dataset prevent

independent replication, and the strong performance may be partly attributable to the CLIP ViT-L/14 backbone's pre-training regime rather than the proposed fine-grained training innovations. The two-stage pipeline is depicted in Fig. 8.

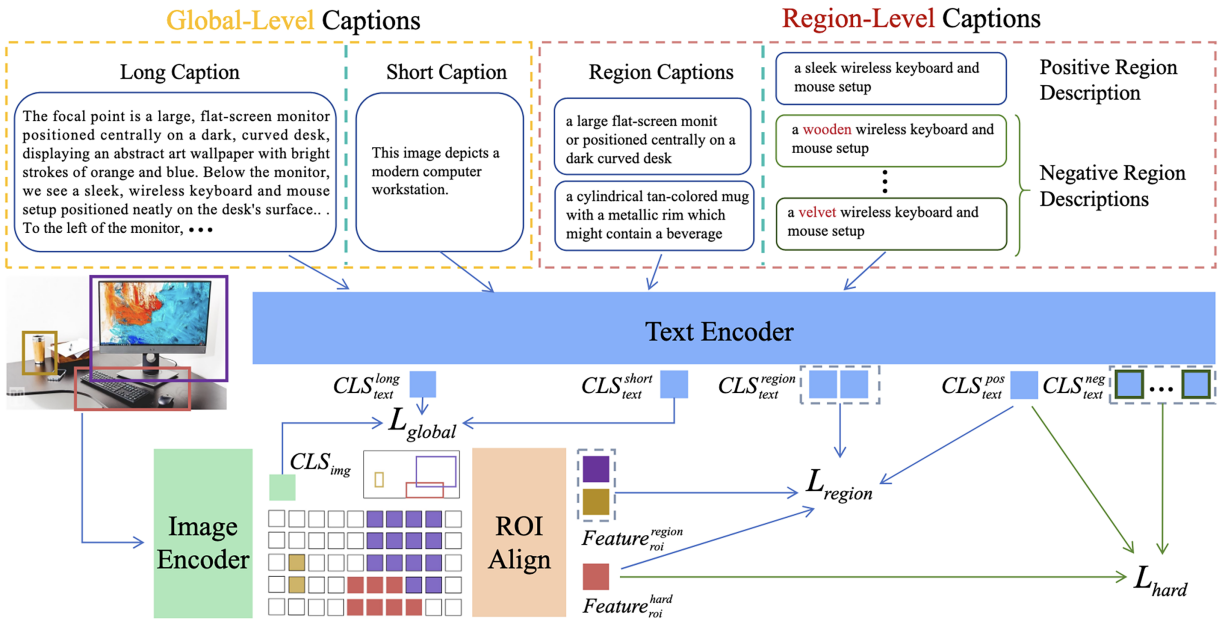

Figure 8: Overview of the FG-CLIP two-stage training pipeline, showing the LMM-recaptioned pre-training stage and the FineHARD regional contrastive stage with RoIAlign and LLM-generated hard negatives [20].

Ref. [33] proposed DetailCLIP (ICLR 2025 workshop), a more modest-scale but methodologically distinctive CLIP extension whose primary evaluation is segmentation rather than fine-grained classification. The framework redirects CLIP's representational energy from global semantics to local spatial structure via attention-guided token removal and patch-level self-distillation, without requiring the massive data or compute of FG-CLIP. A teacher-student pair, both using ViT-B/16, is trained with EMA teacher update (λ increasing from 0.996 to 1.0). The teacher processes two unmasked views; attention values averaged over all heads and layers identify the bottom 50% of patches by attention score, which are removed before the student receives its input. This inverts MAE's random masking strategy: removing low-attention patches rather than masking random patches is principled and directly motivated by the observation that random masking can accidentally remove discriminative patches. The student must: (1) predict patch-level feature distributions matching the teacher (L_{Patch}: KL divergence); (2) align global `[CLS]` token distributions (L_{CLS}: KL divergence); and (3) reconstruct original image pixels (L_{Rec}: MSE via a lightweight decoder). The combined loss is $L_{\text{tot}} = L_{\text{CLS}} + L_{\text{Patch}} + L_{\text{Rec}} + L_{\text{CLIP}}$. DetailCLIP is evaluated primarily on segmentation and zero-shot classification benchmarks rather than standard FGVC benchmarks, limiting direct comparison with the other papers in this survey.

Ref. [34] extended FG-CLIP to a bilingual (English+Chinese) model, FG-CLIP 2 (to appear, ICML 2026), addressing two limitations of the original: its English-only tokenizer and the tendency for contrastive training to collapse text embeddings for attribute-similar captions. FG-CLIP 2 replaces the CLIP text encoder with a multilingual Gemma tokenizer (256K vocabulary) and adopts SigLIP 2 as the base architecture, using a sigmoid binary classification loss rather than InfoNCE's softmax normalisation, which is more stable at the large batch sizes used in vision-language pre-training. Two new training objectives address

text-space discrimination specifically. The Textual Intra-modal Contrastive (TIC) loss computes pairwise cosine similarities between region captions in the batch and treats pairs with similarity above 0.95 as hard text negatives, enforcing separation in the text embedding space independently of vision. The Cross-Modal Margin Ranking (CMR) loss uses adaptive per-negative thresholds computed from the previous training step's similarity distribution, synchronised across all GPUs via all-reduce to estimate the global similarity distribution rather than a mini-batch approximation. Stage 2 training combines five losses: $L = 1.0 \cdot L_{\text{global}} + 0.1 \cdot L_{\text{FGV}} + 0.5 \cdot L_{\text{FGT}} + 0.4 \cdot L_{\text{CMR}} + 0.1 \cdot L_{\text{TIC}}$. Evaluated across 29 datasets and 8 task types, FG-CLIP 2 outperforms FG-CLIP on English fine-grained benchmarks and extends competitive performance to Chinese fine-grained tasks. The key limitation is the proprietary in-house Chinese dataset (12M images) that limits reproducibility of the bilingual component.

8.2 Domain Adaptation and Open-Set Generalisation

Most FGVC methods assume closed-set evaluation: all test-time categories are present in the training set, and the training and test images come from the same domain. Real-world deployments violate both assumptions.

Ref. [24] formulated WDAN as a solution to fine-grained open-set domain adaptation (FG-OSDA). The system aims to align fine-grained features between source and target domains for known categories while preventing unknown-class samples from contaminating the alignment through a weighted adversarial mechanism. A Selective Transformer Encoder (STE) modifies ViT with overlapping patch tokenisation to preserve local fine-grained details. Dual classifiers (C1, C2), trained independently on labeled source data, provide complementary discriminative signals; their disagreement on a target sample is used as a proxy for unknown-class membership. A Weighted Conditional Domain Discriminator (WCDD) performs adversarial domain alignment weighted by $w_i = 1 - P(\text{unknown}|x_i)$, down-weighting or excluding likely-unknown samples from alignment gradients. The paper also introduces the WebStanfordCars FG-OSDA benchmark, an important methodological contribution independent of the WDAN architecture. The weight estimation depends on dual-classifier calibration; if both classifiers agree on a wrong prediction for an unknown-class sample, that sample will be incorrectly aligned.

8.3 Prompt-Based ViT Adaptation for Interpretable Fine-Grained Recognition

CLIP and DINO represent complementary pre-training philosophies: CLIP trains on image-text pairs and produces language-aligned representations, while DINO trains via self-supervised local-global consistency and produces spatially discriminative patch features without linguistic grounding. Prompt-based adaptation of DINO-type models for fine-grained recognition exploits these spatially rich patch features directly.

Prompt learning originated as a parameter-efficient alternative to full fine-tuning of vision-language models, and the fine-grained methods below specialise it to subordinate-category recognition. CoOp replaced CLIP's hand-crafted text prompts with a set of learnable context vectors optimised on a few labelled images per class [35], and CoCoOp made those prompts input-conditional to recover the generalisation to unseen classes that CoOp's static context lost [36]. MaPLe extended prompting to the vision and language branches jointly, coupling the two modalities across successive transformer stages [37]. These methods established prompt learning as the dominant lightweight adaptation strategy for foundation models; the two fine-grained methods that follow inherit this machinery and adapt it to the demands of fine-grained discrimination.

Ref. [38] proposed Prompt-CAM, which adds C learnable class-specific prompt tokens to the last layer of a frozen pre-trained ViT (primarily DINO ViT-B) and uses the resulting class-specific multi-head attention maps directly for both classification and fine-grained trait localisation. The key architectural insight is a deliberate constraint: a single shared scoring vector $w \in \mathbb{R}^D$ (not class-specific) is used across all classes, so that $s[c] = w^T z_N^c$ where z_N^c is the output feature of class c's prompt. Because w cannot encode class-discriminative information in patch value features, the model is architecturally forced to route all class-specificity through the attention weights α^c, which attend to unique image patches (traits) for the true class. Class-agnostic prompts are injected to all earlier layers for backbone adaptation. Training uses only cross-entropy on $s[c]$; the backbone is fully frozen throughout. At inference, the R attention maps from R heads in the last layer reveal up to R distinct class-specific traits. A greedy head masking procedure identifies the minimal set of attention heads necessary for correct classification, ranking traits by diagnostic importance. On CUB-200-2011 using a DINO backbone, Prompt-CAM achieves 73.2% accuracy (vs. 78.6% for linear probing), with a faithfulness advantage on insertion/deletion metrics (insertion 0.61, deletion 0.09) over Grad-CAM, Layer-CAM, and attention rollout. Human participant studies confirm that 60.49% of expert-identified traits are recognisable in Prompt-CAM heatmaps, compared to 39.14% for TesNet and 30.39% for ProtoConcepts. The accuracy-interpretability trade-off is real and honestly reported; images with occluded or unusual-pose traits are the primary failure mode.

Ref. [39] took prompt learning in a generative direction with GCPL (Generative Class Prompt Learning), which conditions a text-to-image diffusion model on few-shot exemplars through learnable class prompts, strengthening the visio-linguistic coupling of the class embeddings rather than only the discriminative boundary. A contrastive extension, CoMPLe, adds an inter-class separation term to the generative optimisation. The method targets the weak fine-grained discrimination of off-the-shelf VLM embeddings and reports improved few-shot recognition over prior prompt-learning baselines; as a diffusion-based generative approach, however, it inherits the sampling cost of its backbone, and the paper appeared as a preprint before its conference version.

8.4 Taxonomy

Fig. 9 organises the seven surveyed methods into three sub-themes: CLIP adaptation for fine-grained vision-language alignment, domain adaptation and open-set generalisation, and prompt-based adaptation. The general prompt-learning lineage (CoOp, CoCoOp, MaPLe) is discussed as background rather than classified as a surveyed FGVC method.

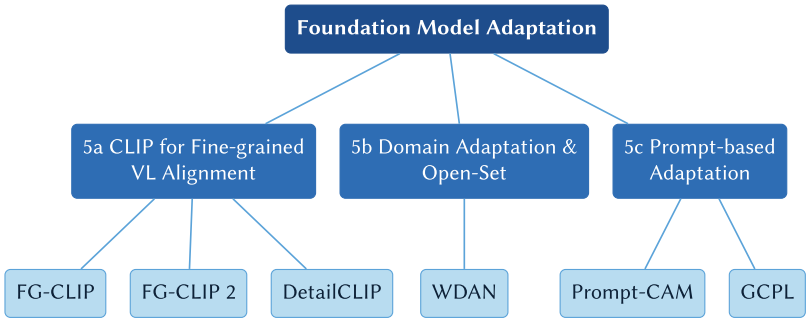

Figure 9: Taxonomy of foundation model adaptation methods (seven surveyed papers across three sub-themes; the general prompt-learning lineage of CoOp, CoCoOp, and MaPLe is discussed as background): FG-CLIP [20], FG-CLIP 2 [34], DetailCLIP [33], WDAN [24], Prompt-CAM [38], GCPL [39].

8.5 Synthesis of Foundation Model Adaptation

FG-CLIP, FG-CLIP 2, and DetailCLIP pursue the same underlying goal, closing CLIP's coarse-global-alignment gap for fine-grained tasks, but resolve it through strategies that are almost perfectly orthogonal. FG-CLIP's response is scale: recaption 1.6 billion image-text pairs with a large multimodal model and add region-level contrastive supervision at scale. FG-CLIP 2 extends this approach in two targeted directions: multilingual reach (Gemma 256K tokenizer, Chinese training data) and text-space discrimination (TIC loss, CMR loss). The TIC loss is conceptually the most interesting new contribution: it enforces that text embeddings for visually distinct captions are separated in the text embedding space independently of any vision signal, addressing a failure mode of contrastive training where attribute-similar captions collapse in the text encoder. DetailCLIP's response is architectural: add patch-level self-distillation and pixel-level reconstruction objectives that redirect the model's representational energy to local spatial structure. FG-CLIP is more of an engineering achievement than a methodological insight; its most important contribution may be the FineHARD dataset construction pipeline, which demonstrates that LLM-generated attribute-modified hard negatives can be produced at scale with 98.9% quality. DetailCLIP is more methodologically interesting from a self-supervised learning perspective: inverting MAE's random masking strategy, removing low-attention patches rather than masking random patches, is principled and directly motivated by the observation that random masking can accidentally remove exactly the discriminative patches the model should learn from.

Prompt-CAM introduces a qualitatively different goal: not better recognition accuracy, but reliable fine-grained interpretability. Its core insight, using a shared (not class-specific) scoring vector to force class identity into attention weights rather than value features, is a rare example of interpretability achieved through architectural constraint rather than post-hoc analysis. The resulting −5.4% accuracy trade-off on CUB is real, but the faithfulness gains (insertion 0.61 vs. 0.52 for Grad-CAM) and human recognition rates (60.49% vs. 39.14% for TesNet) suggest this trade-off is worth accepting for biological trait discovery applications. The greedy head masking procedure for trait ranking is an underexplored direction: automatically identifying the minimal discriminative evidence is valuable for both AI interpretability and scientific hypothesis generation.

A deeper issue running through this theme is what WDAN's results imply about the standard evaluation framework. WDAN demonstrates that competitive fine-grained recognition can be achieved with a comparatively simple ViT backbone under realistic open-set conditions, without massive-scale pre-training. That this is presented as a domain adaptation paper rather than a standalone FGVC method reflects the field's tendency to evaluate methods exclusively in the closed-set, same-domain regime. FG-CLIP generalises across multiple downstream tasks but achieves this through scale rather than through an architecture designed for cross-domain fine-grained generalisation. The theoretical relationship between the scale of pre-training, the specificity of fine-grained supervision, and cross-domain generalisability remains analytically underexplored. Table 6 summarises the methods surveyed in this theme alongside their core problem, method, strength, weakness, and evaluation metric.

Table 6: Reference summary for foundation model adaptation.

Reference	Core Problem	Method	Strength	Weakness	Metric
FG-CLIP [20]	Coarse CLIP captions lack regional detail	Stage 1: 1.6B LMM-recaptioned pairs + global InfoNCE. Stage 2: FineHARD (12M imgs, 40M RoI, 10M LLM hard negatives) + $L_{\text{global}} + 0.1L_{\text{regional}} + 0.5L_{\text{hard}}$	Massive-scale coverage; RoIAlign enables explicit regional alignment; generalises across classification, detection, retrieval	~30 days on 160 NPUs; FineHARD proprietary; LLM hard negatives may introduce label noise	Top-1 accuracy
DetailCLIP [33]	Random masking discards discriminative patch tokens	Teacher-student EMA; attention-guided removal of bottom 50% patches; student $L_{\text{CLS}} + L_{\text{Patch}} + L_{\text{Rec}} + L_{\text{CLIP}}$	Multi-objective SSL captures global semantics and local spatial detail; attention removal more principled than random masking	Not evaluated on standard FGVC benchmarks; noisy early attention may remove discriminative tokens	mIoU/zero-shot accuracy
WDAN [24]	Unknown classes corrupt domain alignment	STE (overlapping-patch ViT) + dual classifiers (disagreement = unknown proxy) + WCDD weighted adversarial alignment	First FG-OSDA benchmark (WebStanford-Cars); weighted alignment prevents negative transfer	Weight estimation fragile when both classifiers agree incorrectly; limited to visually close domain pairs	H-score
FG-CLIP 2 [34]	English-only CLIP; text embeddings collapse for attribute-similar captions	SigLIP 2 + Gemma 256K tokenizer + TIC (cosine sim > 0.95 text hard negatives) + CMR (global adaptive threshold ranking); $L = L_{\text{global}} + 0.1L_{\text{FGV}} + 0.5L_{\text{FGT}} + 0.4L_{\text{CMR}} + 0.1L_{\text{TIC}}$	Bilingual (EN+ZH); TIC enforces text-space discrimination independently of vision; CMR global threshold more accurate than mini-batch	Proprietary Chinese data; 5-loss weight tuning; L_{TIC} requires $O(B^2)$ text similarity computation	Top-1 accuracy
Prompt-CAM [38]	Post-hoc saliency (Grad-CAM) produces class-agnostic blurred heatmaps	C class-specific prompts on frozen DINO ViT-B last layer; shared scoring vector w (not class-specific) forces class identity into attention weights α^c; greedy head masking for trait ranking	Architecturally forced interpretability; faithfulness insertion 0.61 vs. Grad-CAM 0.52; human trait recognition 60.49% vs. 39.14% (TesNet)	−5.4% accuracy vs. linear probing; failure on occluded/unusual-pose traits; large-scale expert annotation not validated	Faithfulness (insertion/deletion) + accuracy

(Continued)

Table 6 (continued)

Reference	Core Problem	Method	Strength	Weakness	Metric
GCPL [39]	Off-the-shelf VLM embeddings discriminate fine-grained classes weakly	Text-to-image diffusion conditioned on few-shot exemplars via learnable class prompts; CoMPLe adds inter-class contrastive term	Strengthens visio-linguistic coupling of class embeddings; improves few-shot recognition over prompt-learning baselines	Diffusion sampling cost; preprint prior to conference version	Few-shot top-1 accuracy

9 Ultra-Fine-Grained Visual Categorization as an Emerging Sub-Field

The methods surveyed above treat ultra-fine-grained visual categorization (ultra-FGVC) as one application among several, with its representative works distributed across four of the five themes: SV-Transformer under region discovery (Section 4.4), CLE-ViT and CLA-Net under metric learning (Section 5.1), and CSDNet under training strategy (Section 6.4). This distribution reflects how each method was proposed, but it obscures a coherent trend: ultra-FGVC is consolidating into a distinct sub-field with its own benchmarks, its own dominant failure modes, and a set of design commitments that do not follow from standard FGVC. This section draws the scattered threads together.

9.1 Problem Definition and Benchmarks

Standard FGVC separates subordinate categories within a genus or a product line, such as bird species or car models. Ultra-FGVC pushes one level finer, to cultivars within a single species, where inter-class appearance is so similar that human experts struggle to assign labels reliably [40]. The reference benchmark is the ultra-fine-grained (UFG) dataset of [40], which contains 47,114 images across 3526 cultivar categories, aggregated into five sub-datasets that recur throughout the ultra-FGVC literature (Table 7). Two properties define the regime. First, data scarcity is extreme: most sub-datasets provide only three to eleven images per category, an order of magnitude fewer than CUB-200-2011. Second, the discriminative signal is not only subtle but frequently non-object: leaf specimens are photographed against variable backgrounds, so a model that latches onto background texture rather than the cultivar-specific foreground fails even when it appears to train successfully.

Table 7: The five commonly used sub-datasets of the ultra-fine-grained (UFG) benchmark [40]. Class counts and splits are cross-confirmed across independent reports; per-category sample counts are approximately three to eleven images on the cotton and soybean cultivar sets.

Sub-Dataset	Classes	Train	Test	Domain
Cotton80	80	240	240	Cotton cultivars
SoyLocal (SoyLoc)	200	600	600	Soybean cultivars
SoyGene	1110	12,763	11,143	Soybean (genetic)
SoyAgeing	198	4950	4950	Soybean (growth stages R1–R6)
SoyGlobal	1938	5814	5814	Soybean (global)

9.2 What Ultra-FGVC Methods Do Differently

The methods that target this regime converge on three commitments that standard FGVC methods do not share. The first is explicit background control. SV-Transformer identifies background patches by their low attention to a global object token and replaces them with distractor patches during training, forcing background invariance; its ablation shows a larger gain on the high-background-variation cultivar sets (Cotton80, SoyCultivar) than on CUB, reaching Cotton80 72.08%, SoyLocal 65.17%, and SoyGene 88.83% (Section 4.4). The second is hardened self-supervision to compensate for scarce labels. CLE-ViT constructs hard positives by shuffling and masking a single image, since ordinary augmented positives are too easy in the low-data regime, yielding a reported +19% on SoyGlobal (Section 5.1); CLA-Net replaces instance-level feature congruence with a softer covariance-manifold (Lie algebra) constraint to avoid catastrophic overfitting on three-to-eleven-image classes. CSDNet attacks scarcity at three levels at once (discriminative-region parsing, a memory-queue contrastive loss, and self-distillation from an augmented branch that is discarded at inference), reaching Cotton80 67.92%, SoyLocal 60.50%, and SoyGene 86.86% while retaining 113.67 FPS single-branch inference (Section 6.4).

Two recent works extend the sub-field along axes the earlier methods left open. Ref. [41] target the adaptation cost directly: rather than fine-tuning the whole backbone, they freeze it and train only an Intermediate Layer Adapter with dual-branch spatial down-sampling between encoder groups, reporting an average accuracy improvement of at least 6.8% in the parameter-efficient setting while using roughly two orders of magnitude fewer trainable parameters and about 30% fewer FLOPs than prior ultra-FGVC methods, evaluated across the ten UFG leaf datasets. Ref. [42] argue that geometric structure, such as leaf vein configuration, is an overlooked cue: their self-supervised network amplifies geometry-relevant detail and encodes the relative polar coordinates of discriminative parts under rotation-invariant constraints, then distils the geometric pathway into the classification branch so inference uses only raw images. On the five UFG benchmarks the preprint reports Cotton80 71.3%, SoyLoc 62.5%, SoyGene 88.7%, SoyAgeing 86.4%, and SoyGlobal 81.2% with a Swin-B backbone; these figures are from an unrefereed preprint and should be read with the caution that Section 10.4 recommends.

9.3 Trajectory

Read together, these methods show ultra-FGVC maturing from a set of hard benchmarks into a sub-field with its own design language. Background suppression, single-image hard-positive construction, geometry-aware self-supervision, and parameter-efficient adaptation are all responses to the same pair of pressures (extreme label scarcity and background-dominated signal) that standard FGVC does not face at the same intensity. The regime also exposes the evaluation problems of Section 10 in sharp relief: backbones range from ResNet-50 to Swin-B, results are reported on overlapping but non-identical subsets of the UFG datasets, and several headline numbers originate in preprints. The regime-separation and evidence-provenance rows of the checklist in Table 8 are therefore especially load-bearing here. The most plausible near-term directions are foundation-model adaptation under cultivar-level scarcity, where the parameter-efficient adapters of [41] meet the vision-language methods of Section 8, and a standardised ultra-FGVC protocol that fixes the backbone and the sub-dataset split so that the sub-field's rapid recent gains can be attributed to mechanism rather than configuration.

Table 8: A Checklist for fair and reproducible FGVC evaluation, distilled from the evaluation weaknesses recurring across the surveyed literature. Each item is a minimum reporting requirement; the rationale column names the confound it controls.

Axis	Requirement	Rationale/Confound Controlled
Backbone parity	Report the backbone family, pretraining dataset, and parameter count; compare methods only against baselines using the same backbone and pretraining.	Isolates the method's contribution from backbone and pretraining-scale advantages (Section 10.2).
Baseline ladder	Report, on a fixed backbone, the frozen-backbone linear probe, full fine-tuning, and the proposed method.	Bounds the method's gain between the trivial floor and the full-fine-tuning ceiling.
Capacity and cost	Report trainable and total parameters, FLOPs, inference forward passes per image, and training budget (device-hours and device type).	Exposes hidden costs such as two-pass inference and large-scale pretraining budgets.
Multi-benchmark	Report CUB-200-2011, Stanford Cars, and FGVC-Aircraft together, not a single dataset.	Prevents dataset-specific tuning from being read as general fine-grained capability (Section 10.3).
Data protocol	State train/test splits, input resolution, and the full augmentation pipeline; use the standard splits unmodified.	Removes split, resolution, and augmentation as uncontrolled sources of gain.
Statistical rigour	Report mean and standard deviation over multiple seeds, not a single run.	Distinguishes a real gain from seed variance.
Localisation vs. classification	Where a method claims better localisation, report a localisation metric (e.g., using CUB part annotations) alongside top-1.	Attributes accuracy gains to localisation or feature quality rather than the full system (Section 10.1).
Regime separation	Report standard-FGVC and ultra-FGVC results separately; never average across the two regimes.	The two regimes have qualitatively different data and difficulty (Section 9).
Evidence provenance	When tabulating state-of-the-art numbers, distinguish peer-reviewed results from preprints.	Prevents unrefereed single-paper results from being read as established consensus.

10 Discussion

10.1 The Inseparability of Region Discovery and Feature-Space Shaping

The most consistent finding across the surveyed literature is that methods which address localisation and feature discrimination jointly outperform those that treat them as sequential sub-problems. TransFG aggregates all-layer attention to identify informative tokens and then imposes contrastive supervision on confusing subcategory pairs; both operations are active simultaneously during training. FEM pairs a weakly supervised saliency selector with a class-centric loss that directly controls the geometry of the feature space

around selected regions. HI2R constructs a hypergraph over attended tokens and uses an inter-category prototype to regularise the feature space. In each case, the two objectives co-supervise each other: a more discriminative feature space produces better attention maps, which produce better region masks, which produce better features. Methods that separate the two, for example CAM-based localisation feeding into a standard cross-entropy classifier as in FGVC-CAM, hit a ceiling defined by the quality of the upstream locator, which is itself trained without feedback from the downstream feature quality.

This inseparability has a practical implication for the field's evaluation culture. Reporting top-1 accuracy on CUB-200-2011 does not distinguish between accuracy gains that come from better localisation, better feature-space shaping, or simply a stronger backbone. The lack of a standard localisation evaluation protocol means that improvements are routinely attributed to the full system when they may be largely explained by a single component. Several papers in this survey do not include ablation studies adequate to disentangle these contributions.

10.2 Backbone Scale as an Uncontrolled Variable

A recurring methodological problem in this literature is the conflation of method-level improvements with backbone-level advantages. Comparing PCA-Net (ResNet-101) to FEM (ViT-B/16 pretrained on ImageNet-21K) on CUB-200-2011 is not a fair comparison of region discovery strategies; it is partially a comparison of what ResNet-101 and ViT-B/16-21K encode before any task-specific training begins. FG-CLIP's top performance across benchmarks is partly attributable to its CLIP ViT-L/14 backbone, which was trained on 400 million image-text pairs, a pretraining regime qualitatively different from ImageNet-21K. The field has not established a norm of reporting results under backbone-controlled conditions, and the absence of such a norm makes it difficult to isolate what the proposed method actually contributes. DDA-Trans is the most egregious case: the choice of 11 layers instead of the standard 12 gives the method a FLOPs advantage not attributable to the dual-path attention mechanism.

A more rigorous evaluation culture would require, at minimum, a comparison against a strong baseline using the same backbone, ideally the frozen backbone with a linear head, the backbone with full fine-tuning, and the backbone with the proposed method, on all three standard benchmarks. Several papers in this survey report only the full-method result on a single benchmark, making any inference about the method's actual contribution unreliable.

10.3 Cross-Dataset Generalisation

The evaluation practice across this literature is almost exclusively within-dataset: train and test on CUB-200-2011, report accuracy, compare to prior work. No paper in this survey systematically evaluates whether a method trained on CUB-200-2011 transfers to Stanford Cars or FGVC-Aircraft, despite the fact that these benchmarks test qualitatively different forms of fine-grained discrimination. This matters because the methods surveyed make implicit claims about fine-grained recognition in general, not specifically about bird species. A region discovery method that learns to attend to plumage markings may not generalise to badge-level detail on car images, and a frequency-domain architecture designed to capture texture in feathers may not capture the geometric subtleties that distinguish aircraft variants.

The one paper that partially addresses this is FG-CLIP, which reports results across classification, detection, and retrieval tasks and demonstrates generalisation across multiple FGVC domains. That it does so through massive-scale training rather than a principled cross-domain architecture design is consistent with the hypothesis that generalisation in FGVC is currently purchased through scale rather than architectural insight. A more productive research agenda would investigate what structural properties of a method predict

cross-dataset generalisation, and use this as an additional evaluation criterion alongside within-dataset top-1 accuracy.

10.4 A Checklist for Fair FGVC Evaluation

The evaluation weaknesses catalogued above (uncontrolled backbones, single-benchmark reporting, conflated localisation and feature-space contributions, and mixed standard and ultra-FGVC results) recur across the surveyed literature and are individually avoidable. We distil them into the checklist in Table 8, intended as a concrete reporting protocol for authors and a review aid for referees. The checklist does not prescribe a method; it prescribes what must be reported for a top-1 accuracy comparison to carry a defensible causal claim about the proposed mechanism. Most items cost nothing beyond disclosure, and several (the backbone-parity baseline ladder, the standard-vs.-ultra regime split) directly address the confounds that Sections 10.2 and 10.3 identify as the most damaging to cross-paper comparability.

11 Future Directions

11.1 Occlusion-Robust Fine-Grained Recognition

Occlusion is the most practically significant failure mode in FGVC and the least studied. Of the thirty papers surveyed, only [11] directly addresses it, and even that work relies on a cyclical attention mechanism whose convergence properties and optimal iteration count are not characterised. The standard benchmarks contain mostly unoccluded, well-lit, centred subjects photographed under controlled conditions. Real-world FGVC deployments, including wildlife monitoring, vehicle identification, and aerial surveillance, routinely involve partial occlusion, motion blur, and variable pose. Developing evaluation protocols that specifically test occlusion robustness, and methods that address it architecturally rather than as an afterthought, is a high-priority direction.

11.2 Unified Localisation and Feature-Space Evaluation

The field would benefit from an evaluation protocol that separately measures localisation quality and feature-space discriminability rather than conflating both into a single top-1 accuracy number. CUB-200-2011 provides part location annotations that are almost never used in recent work; reintegrating them into the evaluation framework would allow researchers to isolate whether performance gains come from better localisation, better feature extraction, or both. Analogous to the way that object detection benchmarks decompose performance into localisation (IoU) and classification (mAP), an FGVC evaluation standard that reports part localisation accuracy alongside category accuracy would substantially improve the field's ability to make targeted methodological progress.

11.3 Theoretical Foundations

The theoretical basis for FGVC methods is almost entirely absent from the surveyed literature. It is not known, analytically, what properties a feature space must have to support fine-grained discrimination; what class of attention mechanisms is sufficient to reliably localise discriminative regions; or what the relationship is between the rank of a weight update and the fine-grained discriminability of the resulting representation. Addressing this gap would provide principled guidance for architecture design rather than the current empirical trial-and-error culture.

12 Conclusion

This review has surveyed thirty methods spanning 2021 to 2026 that address the core challenge of fine-grained visual categorisation: the discriminative signal is spatially concentrated, visually subtle, and easily destroyed by the preprocessing strategies that serve coarser recognition tasks. Organising these methods under five analytical themes has revealed that the field is not best characterised as a sequence of independent technical innovations, but as a sustained negotiation between two inseparable sub-problems (where to look and how to represent what is found).

The most important empirical regularity across the literature is that methods which address region discovery and feature-space shaping jointly achieve higher and more consistent accuracy than those that treat them as sequential stages. TransFG, FEM, and HI2R each demonstrate that co-supervising attention and representation learning produces qualitatively different results from applying them in sequence. This convergence, arrived at independently by multiple groups, constitutes the strongest result the review can offer: joint supervision of localisation and feature geometry is not one design choice among many but the structurally superior approach.

A second and underappreciated regularity is the systematic undercontrol of backbone strength as an experimental variable. The largest performance differences across the surveyed papers, for instance between PCA-Net at 88.2% and FG-CLIP at 93.8% on CUB-200-2011, are not primarily explained by methodological innovation; they reflect the difference between ResNet-101 trained on ImageNet-1K and CLIP ViT-L/14 trained on 1.6 billion image-text pairs. Until the field establishes a norm of comparing methods on a fixed backbone, the attribution of accuracy gains to specific contributions will remain unreliable.

The principal open problems, identified but not resolved by the surveyed literature, are three. First, occlusion robustness: only one method (SRL-FGVC) directly targets heavy occlusion, and even that work does not characterise the convergence properties of its cyclical mechanism. Second, cross-dataset generalisation: no paper evaluates whether a model trained on CUB-200-2011 transfers to Stanford Cars or FGVC-Aircraft without re-training, leaving the field's generalisation claims empirically untested. Third, theoretical foundations: the conditions under which an attention mechanism will reliably localise discriminative regions, and the relationship between low-rank weight updates and fine-grained discriminability, are not analytically characterised. Addressing these three problems would transform FGVC from an empirical benchmark competition into a field with principled, predictive design guidelines. Appendix A provides the full five-theme taxonomy diagram, and Appendix B (Tables A1–A4) consolidates all thirty surveyed methods in a unified reference summary.

Acknowledgement: The authors thank colleagues at the University for Development Studies and at Kwame Nkrumah University of Science and Technology for ongoing discussions on fine-grained recognition that informed several of the framings adopted in this review.

Funding Statement: The authors received no specific funding for this study.

Author Contributions: The authors confirm contribution to the paper as follows: conceptualisation, Richard Adusei and Gaddafi Abdul-Salaam; methodology (literature search and inclusion criteria), Richard Adusei; formal analysis (paper-level summarisation and taxonomy construction), Richard Adusei and Gaddafi Abdul-Salaam; investigation (reading and synthesis of the surveyed literature), Richard Adusei; writing—original draft preparation, Richard Adusei; writing—review and editing, Richard Adusei and Gaddafi Abdul-Salaam; visualisation (taxonomy diagrams and reference tables), Richard Adusei; supervision, Gaddafi Abdul-Salaam; project administration, Richard Adusei. All authors reviewed and approved the final version of the manuscript.

Availability of Data and Materials: Not applicable. This article does not involve data availability, and this section is not applicable, as the work is a narrative review and all surveyed material is cited in the reference list.

Ethics Approval: Not applicable. This study did not involve human or animal subjects.

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

Appendix A Full Survey Taxonomy

Fig. A1 shows all five themes, their sub-themes, and the paper-level classification of the surveyed methods. Papers that contribute to multiple themes are listed under the theme in which their most distinctive contribution falls. The diagram is rendered on a landscape page to accommodate all levels of the hierarchy.

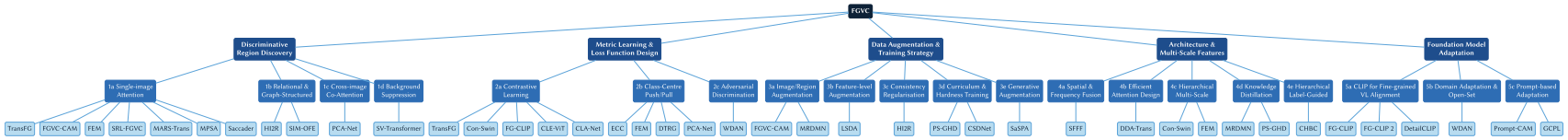

Figure A1: Full taxonomy of the surveyed FGVC literature, showing all five themes, seventeen sub-themes, and the paper-level classification of the methods assigned to the thematic axes. Papers appearing in multiple sub-themes reflect cross-cutting contributions; the cross-cutting ultra-FGVC methods and benchmark are treated separately in Section 9. Methods shown: TransFG [1], FGVC-CAM [10], FEM [2], SRL-FGVC [11], MARS-Trans [12], MPSA [13], Saccader [14], HI2R [15], SIM-OFE [16], PCA-Net [17], SV-Transformer [18], Con-Swin [19], FG-CLIP [20], CLE-ViT [22], CLA-Net [23], ECC [6], DTRG [21], WDAN [24], MRDMN [25], LSDA [26], PS-GHD [27], CSDNet [28], SaSPA [29], SFFF [30], DDA-Trans [31], CHBC [32], FG-CLIP 2 [34], DetailCLIP [33], Prompt-CAM [38], GCPL [39].

Appendix B General Summary of Surveyed Papers

The tables below provide a consolidated reference for all papers covered in this survey, listing the method name, primary theme, backbone, evaluated datasets, best reported accuracy, and core contribution.

Table A1: General summary of the surveyed FGVC papers, Part 1 of 4: 2021–2024. Papers in multiple themes are listed under their primary theme. Accuracy figures are top-1 on CUB-200-2011 unless otherwise noted.

Method	Year	Reference	Theme	Backbone	Best Result	Cost/Complexity	Core Contribution
TransFG	2021	[1]	Discriminative Region Discovery; Metric Learning	ViT-B/16 (IN-21K)	CUB 91.7%	$O(B^2)$ contrastive; 12-layer rollout	Attention rollout PSM selects top-K discriminative tokens; pairwise contrastive loss on selected features
PCA-Net	2021	[17]	Discriminative Region Discovery	ResNet-101	CUB 88.2%	—	Progressive cross-image co-attention with iterative erasure discovers complementary discriminative channels
FGVC-CAM	2022	[10]	Discriminative Region Discovery; Data Augmentation	ResNet-50	CUB 88.5%	No extra params; 4-branch loss	CAM-guided Region Crop, Drop, and Mix augmentation preserves discriminative cues
WDAN	2023	[24]	Metric Learning; Foundation Model	ViT (STE)	H-score	—	Weighted adversarial alignment for fine-grained open-set domain adaptation; WebStanfordCars benchmark
SFFF	2023	[30]	Architecture	ResNet (dual-branch)	CUB 90.3%	Dual-branch doubles FLOPs	Spatial-frequency dual-branch CNN with importance-sparsity channel selection and RSEF exchange
MRDMN	2023	[25]	Data Augmentation; Architecture	ResNet-101	CUB 91.3%	Complex 3-stage training	CAM-guided discriminative region Mixup with area-proportional soft labels; multi-resolution distillation

(Continued)

Table A1 (continued)

Method	Year	Reference	Theme	Backbone	Best Result	Cost/Complexity	Core Contribution
HI2R	2024	[15]	Discriminative Region Discovery; Data Augmentation	ViT-B/16	CUB 92.5%	Concept cache adds memory	Hypergraph-guided intra- and inter-category relation modelling with RSC symmetry regularisation
ECC	2024	[6]	Metric Learning	ResNet-50/ViT	CUB 91.6%	No architectural changes	Most-Confusing-Class push-pull MCC loss with soft label generation from class-centre distributions
LSDA	2024	[26]	Data Augmentation	ViT-B/16	CUB 91.6%	˜3x training cost (meta-learning)	Feature-level implicit augmentation via learned per-class covariance with meta-learning optimisation

Table A2: General summary of the surveyed FGVC papers, Part 2 of 4: 2024–2026 (continued from Table A1).

Method	Year	Reference	Theme	Backbone	Best Result	Cost/ Complexity	Core Contribution
DDA-Trans	2024	[31]	Architecture	ViT (11-layer)	CUB 91.5%	Linear attention; 11-layer ViT	Linear-complexity dual-path attention decomposing positional and semantic dependencies via centre tokens
PS-GHD	2024	[27]	Data Augmentation; Architecture	ViT-B/16 (IN-21K)	CUB 91.9%	˜3x training time (3-stage)	Three-stage progressive hardness distillation with hard-class relabelling and stage-adaptive loss switching
FG-CLIP	2025	[20]	Metric Learning; Foundation Model	CLIP ViT-L/14	CUB 93.8%	˜30 days on 160 NPUs	Massive-scale CLIP fine-tuning with regional RoIAlign supervision and LLM attribute-modified hard negatives

(Continued)

Table A2 (continued)

Method	Year	Reference	Theme	Backbone	Best Result	Cost/ Complexity	Core Contribution
DetailCLIP	2025	[33]	Foundation Model	ViT-B/16	mIoU 48.1%	—	Attention-guided low-saliency token removal with patch-level self-distillation and pixel reconstruction
Con-Swin	2025	[19]	Metric Learning; Architecture	Swin-B	CUB 91.9%	No extra modules or losses	Projection-head-decoupled contrastive loss on four-stage Swin concatenation prevents backbone distortion
FEM	2026	[2]	Discriminative Region Discovery; Metric Learning; Architecture	ViT-B/16 (IN-21K)	CUB 92.6%	Extra GCN overhead	Plug-and-play WSS + dynamic GCN + Class-Centric Loss (Pearson) jointly shapes region selection and feature geometry
SRL-FGVC	2026	[11]	Discriminative Region Discovery; Metric Learning	ViT-B/16	CUB 92.3%	—	Cyclical MIAF/SAMG attention mining under occlusion with cooperative auxiliary branch and consistency loss

Table A3: General summary of the surveyed FGVC papers, Part 3 of 4: additional methods 2021–2024 (continued from Table A2).

Method	Year	Reference	Theme	Backbone	Best Result	Cost/Complexity	Core Contribution
DTRG	2022	[21]	Metric Learning (Class-Centre)	ResNet-50	CUB 90.1%	No architectural changes	Dynamic cosine graph over EMA class centres with GSL Frobenius loss and OCL shapes global inter-class structure
CLE-ViT	2023	[22]	Metric Learning (Contrastive)	Swin-B	SoyGlobal +19%	—	Self-supervised shuffle + mask hard positive construction with triplet loss drives ultra-FGVC intra-class invariance
MARS-Trans	2025	[12]	Discriminative Region Discovery	ViT-B/16	CUB 91.9%	No new params; 1 scalar/layer	MAM head-deviation weighting + RSM patch boundary supplement + AAM logit correction; three orthogonal ViT improvements

(Continued)

Table A3 (continued)

Method	Year	Reference	Theme	Backbone	Best Result	Cost/Complexity	Core Contribution
MPSA	2024	[13]	Discriminative Region Discovery	Swin-B	CUB 92.8%	4 PSA stages; memory-intensive	Differentiable PSA sampling + MPR multi-stage retrospect + Grad-CAM loss jointly localises and supervises parts
SIM-OFE	2024	[16]	Discriminative Region Discovery (Graph)	ViT-B/16	CUB 92.3%	—	Polar-coordinate GCN captures angular inter-part geometry (+0.4% over Euclidean); OFE FPN multi-scale fusion
CSDNet	2024	[28]	Data Augmentation (Curriculum)	ResNet-50	Cotton80 67.92%	113.67 FPS single-branch inference	SSDP + DDL + SSDT three-level curriculum for ultra-FGVC; 113.67 FPS inference; avoids overfitting on 3–11 samples/class

Table A4: General summary of the surveyed FGVC papers, Part 4 of 4: additional methods 2024–2026 (continued from Table A3).

Method	Year	Reference	Theme	Backbone	Best Result	Cost/ Complexity	Core Contribution
CLA-Net	2025	[23]	Metric Learning (Contrastive)	ResNet-50	Cotton80 62.08%	Lie module discarded at inference	Lie algebra tangent-space contrastive learning on SPD covariance manifold; soft constraint prevents ultra-FGVC overfitting
SaSPA	2024	[29]	Data Augmentation (Generative)	ControlNet + BLIP-Diffusion	Cars 93.8%	—	ControlNet edge-map + BLIP-diffusion subject reference synthesises new training images; α scales inversely with data size
Saccader	2025	[14]	Discriminative Region Discovery	MPSA-based	EU-Moths 97.7%	1 peripheral + 4 fixation passes	NMS priority-map fixation sampling prevents patch clustering; weight-shared re-encoder for sequential saccadic attention
CHBC	2025	[32]	Architecture (Hierarchical Label)	ResNet-50	CUB 90.4% (wt.)	—	Gram-Schmidt orthogonal multi-granularity features + JS-divergence bidirectional consistency across taxonomy levels

(Continued)

Table A4 (continued)

Method	Year	Reference	Theme	Backbone	Best Result	Cost/ Complexity	Core Contribution
FG-CLIP 2	2025	[34]	Foundation Model	SigLIP 2 + Gemma	29-dataset eval	$O(B^2)$ text sim; proprietary data	TIC text-space hard negative loss + CMR global adaptive ranking extends FG-CLIP to bilingual EN + ZH
Prompt-CAM	2025	[38]	Foundation Model (Interpretable)	DINO ViT-B	CUB 73.2% (faithful-ness)	Frozen backbone; C prompt tokens	Class-specific prompts on frozen ViT with shared scoring vector force class identity into attention for trait localisation
SV-Transformer	2026	[18]	Discriminative Region Discovery (BG)	Swin-B	Cotton80 72.08%	—	OFE global token selection + BPL background-patch distractor swap suppresses background for ultra-FGVC
GCPL	2024	[39]	Foundation Model (Prompt-based)	VLM + diffusion	Few-shot FGVC	Diffusion sampling cost	Generative class prompt learning conditions text-to-image diffusion on few-shot exemplars; CoMPLe adds inter-class contrastive term
DILA	2024	[41]	Ultra-FGVC (PE adapter)	ViT (frozen)	+6.8% avg (PE)	~123x fewer params; ~30% fewer FLOPs	Intermediate Layer Adapter with dual-branch spatial down-sampling; ~123x fewer trainable params, ~30% fewer FLOPs
GAEor	2026	[42]	Ultra-FGVC (geometric SSL)	Swin-B	Cotton80 71.3% (pre.)	Geo path distilled at inference	Geometry-guided self-supervision encodes polar coordinates of discriminative parts, distilled into the classification branch

References

1. He J, Chen JN, Liu S, Kortylewski A, Yang C, Bai Y, et al. TransFG: a transformer architecture for fine-grained recognition. Proc AAAI Conf Artif Intell. 2022;36:852–60.
2. Wang D, He X, Lyu S, Tian W, Lu Y. Feature enhancement module based on class-centric loss for fine-grained visual classification. IEEE Trans Neural Netw Learn Syste. 2026;37(2):893–906. doi:10.1109/TNNLS.2025.3613791.
3. Wei XS, Song YZ, Mac Aodha O, Wu J, Peng Y, Tang J, et al. Fine-grained image analysis with deep learning: a survey. IEEE Trans Pattern Anal Mach Intel. 2022;44(12):8927–48. doi:10.1109/tpami.2021.3126648.
4. Zhang Y, Chen W, Zang Y. Fine-grained vision categorization with vision transformer: a survey. In: Proceedings of the 2022 IEEE 8th International Conference on Computer and Communications (ICCC); 2022 Dec 9–12; Chengdu, China. p. 1910–5.
5. Xie Y, Gong Q, Luan X, Yan J, Zhang J. A survey of fine-grained visual categorization based on deep learning. J Syst Eng Electro. 2024;35(6):1337–56. doi:10.23919/jsee.2022.000155.
6. Yao H, Miao Q, Zhao P, Li C, Li X, Feng G, et al. Exploration of class center for fine-grained visual classification. IEEE Trans Circ Syst Video Technol. 2024;34(10):9954–66. doi:10.1109/tcsvt.2024.3406443.

7. Wah C, Branson S, Welinder P, Perona P, Belongie S. The caltech-UCSD birds-200-2011 dataset. Pasadena, CA, USA: California Institute of Technology; 2011. Report No.: CNS-TR-2011-001.
8. Krause J, Stark M, Deng J, Fei-Fei L. 3D object representations for fine-grained categorization. In: Proceedings of the 2013 IEEE International Conference on Computer Vision Workshops; 2013 Dec 2–8; Sydney, NSW, Australia. p. 554–61.
9. Maji S, Rahtu E, Kannala J, Blaschko M, Vedaldi A. Fine-grained visual classification of aircraft. arXiv:1306.5151. 2013.
10. He J, Yang B. Fine-grained image recognition method using discriminative region-based data augmentation. Front Comput Intell Syst. 2022;1(3):8–12. doi:10.54097/fcis.v1i3.1973.
11. Kang B, Liang D, Chen D, Ding T, Wei M. Robust fine-grained visual categorization via cyclical attention. IEEE Trans Neural Netw Learn Syst. 2026;37(1):482–96. doi:10.1109/tnnls.2025.3608560.
12. Mei A, Huo H, Xu J, Xu N. Multistage attention region supplement transformer for fine-grained visual categorization. The Visual Comput. 2025;41(3):1873–89. doi:10.21203/rs.3.rs-3845719/v1.
13. Wang J, Xu Q, Jiang B, Luo B, Tang J. Multi-granularity part sampling attention for fine-grained visual classification. IEEE Trans Image Process. 2024;33:4529–42. doi:10.1109/tip.2024.3441813.
14. Schmidt J, Stober S, Denzler J, Bodesheim P. Saccadic vision for fine-grained visual classification. arXiv:2509.15688. 2025.
15. Chen L, Wang Q, Li Z, Yin Y. Hypergraph-guided intra- and inter-category relation modeling for fine-grained visual recognition. In: Proceedings of the 32nd ACM International Conference on Multimedia; 2024 Oct 28–Nov 1; Melbourne, VIC, Australia. p. 8043–52.
16. Sun H, He X, Xu J, Peng Y. SIM-OFE: structure information mining and object-aware feature enhancement for fine-grained visual categorization. IEEE Trans Image Process. 2024;33:5312–26.
17. Zhang T, Chang D, Ma Z, Guo J. Progressive co-attention network for fine-grained visual classification. In: Proceedings of the 2021 International Conference on Visual Communications and Image Processing (VCIP); 2021 Dec 5–8; Munich, Germany. p. 1–5.
18. Jiang X, Fang Z, Shen F, Gao J, Li Z. Progressive feature encoding with background perturbation learning for ultra-fine-grained visual categorization. IEEE Trans Image Process. 2026;35:585–98. doi:10.1109/tip.2026.3651956.
19. Zhang P, Wang X, Con-Swin ZX. Con-Swin: contrastive swin-transformer for fine-grained visual classification. In: Proceedings of the 2025 International Symposium on Machine Learning and Media Computing (MLMC); 2025 Jul 28–30; Harbin, China. p. 1–6.
20. Xie C, Wang B, Kong F, Li J, Liang D, Zhang G, et al. FG-CLIP: fine-grained visual and textual alignment. In: Proceedings of the 42nd International Conference on Machine Learning (ICML); 2025 Jan 7–Jul 19; Vancouver, BC, Canada. p. 68777–93.
21. Liu K, Chen K, Jia K. Convolutional fine-grained classification with self-supervised target relation regularization. IEEE Trans Image Process. 2022;31:5570–84. doi:10.1109/tip.2022.3197931.
22. Yu X, Wang J, Gao Y. CLE-ViT: contrastive learning encoded transformer for ultra-fine-grained visual categorization. In: Proceedings of the Thirty-Second International Joint Conference on Artificial Intelligence; 2023 Aug 19–25; Macao, China. p. 4531–9.
23. Yu X, Pan Z, Zhao Y, Zhang Q, Gao Y. Contrastive lie algebra learning for ultra-fine-grained visual categorization. In: Proceedings of the 33rd ACM International Conference on Multimedia; 2025 Oct 27–31; Dublin, Ireland. p. 3242–50.
24. Li J, Yang L, Wang Q, Hu Q. WDAN: a weighted discriminative adversarial network with dual classifiers for fine-grained open-set domain adaptation. IEEE Trans Circ Syst Video Technol. 2023;33(9):5133–47.
25. Xu K, Lai R, Gu L, Li Y. Multiresolution discriminative mixup network for fine-grained visual categorization. IEEE Trans Neural Netw Learn Syst. 2023;34(7):3488–500. doi:10.1109/tnnls.2021.3112768.
26. Pu Y, Han Y, Wang Y, Feng J, Deng C, Huang G. Fine-grained recognition with learnable semantic data augmentation. IEEE Trans Image Process. 2024;33:3130–44. doi:10.1109/tip.2024.3364500.

27. Wang Y, Guo W, Xu SX, Yuan S. PS-GHD: progressive self-guided hardness distillation for fine-grained visual classification. In: Proceedings of the 2024 International Joint Conference on Neural Networks (IJCNN); 2024 Jun 30–Jul 5; Yokohama, Japan. p. 1–8.
28. Fang Z, Jiang X, Tang H, Li Z. Learning contrastive self-distillation for ultra-fine-grained visual categorization targeting limited samples. IEEE Trans Circ Syst Video Technol. 2024;34(8):7135–48. doi:10.1109/tcsvt.2024.3370731.
29. Michaeli E, Fried O. Advancing fine-grained classification by structure and subject preserving augmentation. In: Advances in neural information processing systems (NeurIPS). Vol. 37, Red Hook, NY, USA: Curran Associates, Inc.; 2024. p. 22316–49.
30. Wang M, Zhao P, Lu X, Min F, Wang X. Fine-grained visual categorization: a spatial-frequency feature fusion perspective. IEEE Trans Circ Syst Video Technol. 2023;33(6):2798–812.
31. Cui S, Hui B. Dual-dependency attention transformer for fine-grained visual classification. Sensors. 2024;24(7):2337. doi:10.3390/s24072337.
32. Gao P, Liang Y, Song Y, Yang Z. Cross-hierarchical bidirectional consistency learning for fine-grained visual classification. arXiv:2504.13608. 2025.
33. Monsefi AK, Sailaja KP, Alilooee A, Lim SN, Ramnath R. DetailCLIP: detail-oriented CLIP for fine-grained tasks. arXiv:2409.06809. 2025.
34. Xie C, Wang B, Kong F, Li J, Liang D, Ao J, et al. FG-CLIP 2: a bilingual fine-grained vision-language alignment model. arXiv:2510.10921. 2025.
35. Zhou K, Yang J, Loy CC, Liu Z. Learning to prompt for vision-language models. Int J Comput Vis. 2022;130(9):2337–48. doi:10.1007/s11263-022-01653-1.
36. Zhou K, Yang J, Loy CC, Liu Z. Conditional prompt learning for vision-language models. In: Proceedings of the 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR); 2022 Jun 18–24; New Orleans, LA, USA. p. 16795–804.
37. Khattak MU, Rasheed H, Maaz M, Khan S, Khan FS. MaPLe: multi-modal prompt learning. In: Proceedings of the 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR); 2023 Jun 17–24; Vancouver, BC, Canada. p. 19113–22.
38. Chowdhury A, Paul D, Mai Z, Gu J, Zhang Z, Mehrab KS, et al. Prompt-CAM: making vision transformers interpretable for fine-grained analysis. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR); 2025 Jun 10–17; Nashville, TN, USA. p. 4375–85.
39. Chattopadhyay S, Biswas S, Vivoli E, Lladós J. Towards generative class prompt learning for fine-grained visual recognition. arXiv:2409.01835. 2024.
40. Yu X, Zhao Y, Gao Y, Yuan X, Xiong S. Benchmark platform for ultra-fine-grained visual categorization beyond human performance. In: Proceedings of the 2021 IEEE/CVF International Conference on Computer Vision (ICCV); 2021 Oct 10–17; Montreal, QC, Canada. p. 10285–95.
41. Rios EA, Oyerinde F, Hu MC, Lai BC. Down-sampling inter-layer adapter for parameter and computation efficient ultra-fine-grained image recognition. In: European Conference on Computer Vision (ECCV) Workshops. Cham, Switzerland: Springer; 2024.
42. Wang S, Luo Y, Wang Z, Li H, Huang Z, Baktashmotlagh M. Geometry-guided self-supervision for ultra-fine-grained recognition with limited data. arXiv:2604.19345. 2026.