Free-Grained Hierarchical Visual Recognition

Seulki Park[1] Zilin Wang[1] Stella X. Yu[1,2]
[1]University of Michigan [2]UC Berkeley
{seulki,zilinwan,stellayu}@umich.edu

Abstract

Hierarchical image recognition aims to predict labels across a semantic taxonomy, typically assuming fine-grained annotations for every image. However, real-world supervision may appear at any level: a distant bird may only be labeled as "Bird", while a clear image allows "Bald eagle". To reflect this reality, we introduce free-grained hierarchical recognition, where training labels can appear at any level of a taxonomy, requiring consistent predictions under partial and mixed supervision. We construct benchmark datasets with varying label granularity and show that existing hierarchical methods degrade significantly in this setting. To address this, we propose simple yet effective approaches that leverage 1) semantic guidance from vision–language models and 2) visual structure through semi-supervised learning. Finally, we study free-grained inference, where the model adaptively selects prediction depth, enabling reliable coarse predictions when fine-grained ones are uncertain. Together, our task, datasets, and methods provide a practical step toward hierarchical recognition in real-world scenarios[1].

1. Introduction

Hierarchical classification [5, 7, 16, 26] predicts a semantic tree of labels (*Bird* → *Bird of prey* → *Bald eagle*), capturing categories from broad to specific. This richer output supports flexible use: An expert may seek *Bald eagle*, while a general user may only need *Bird*. Moreover, predicting the full hierarchy improves robustness and scalability, encouraging models to generalize across levels, and can naturally support extensions like adding new parent or child classes.

For hierarchical classification, existing methods [5, 41] assume *complete supervision*, where every training image is annotated at all taxonomy levels (Fig. 1,3). However, in practice, annotations often exhibit *mixed-granularity* labels [18, 23], i.e., labels at different taxonomy levels, due to factors such as 1) limited visual detail, 2) annotation cost and expertise, and 3) evolving annotation protocols. For example, a distant image or non-expert annotator may assign coarse labels (e.g., *Bird*, *Dog*), whereas a close-up view or expert annotation enables fine-grained labels (e.g., *Bald eagle*, *German shepherd*, Fig. 1,4).

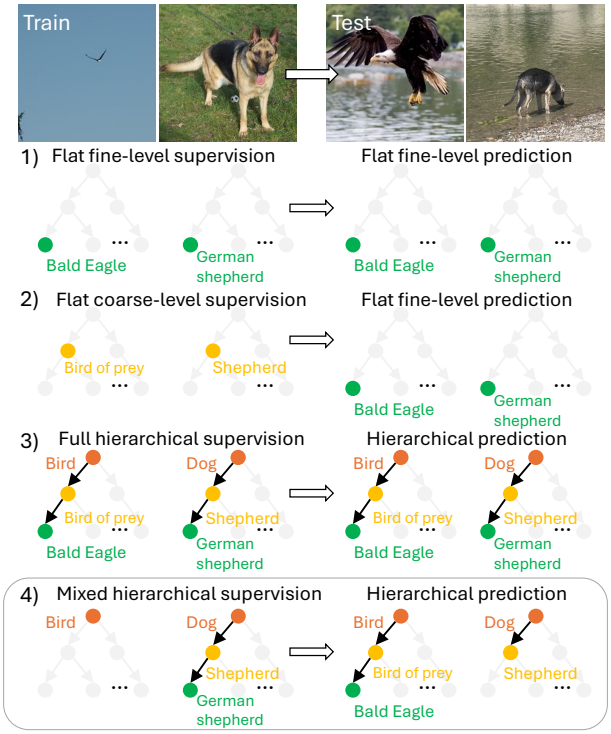


Figure 1. **Free-grained recognition differs from existing tasks in that label granularity can vary during both training and inference. 1) Flat fine-grained recognition**: fine labels → fine predictions (fully supervised) [46]. **2) Flat weakly supervised recognition**: coarse labels → fine predictions [13]. Both 1) and 2) operate on a single flat level without modeling cross-level relations. **3) Hierarchical recognition**: full hierarchy → full hierarchy, requiring complete annotations at all levels [26]. **4) Free-grained recognition**: labels may appear at different levels during training (e.g., a distant bird as *Bird*, a close-up dog as *German shepherd*). At inference, the model predicts the deepest reliable label based on confidence (e.g., a clear bird as *Bald eagle*, but a side-view black dog with ambiguous details as *Shepherd*). This makes the task more challenging, requiring learning semantic and visual relations from mixed and incomplete supervision for consistent hierarchical prediction.

[1]Our dataset and code is available at FreeGrainLearning.

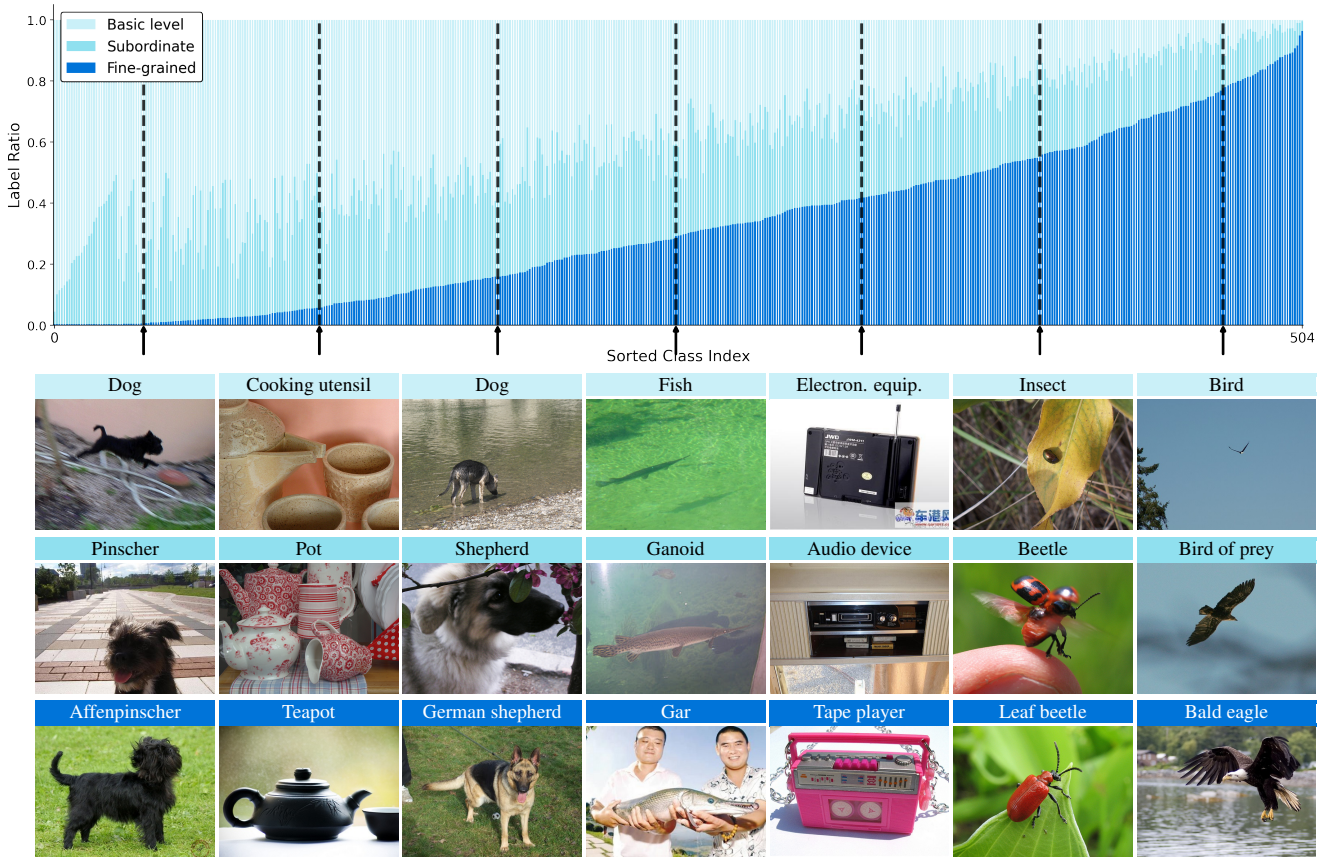


Figure 2. **Our ImageNet-F reflects realistic mixed-granularity supervision, exhibiting both long-tailed fine-grained labels and visual ambiguity. Top:** Distribution of label depth (basic, subordinate, fine) across classes. Fine-grained labels are highly imbalanced, forming a long-tailed pattern where some classes retain many fine labels while others have few. **Bottom:** Examples illustrating confidence-based label assignment using foundation models. (Last column) A distant bird is labeled at the basic level (Bird); a mid-range instance at the subordinate level (Bird of prey); and a clear close-up at the fine-grained level (Bald eagle).

To capture this reality, we propose **free-grained hierarchical visual recognition**, where supervision is free to vary in granularity: training labels may appear at any level of a fixed taxonomy (e.g., *Bird*, *Bald eagle*). At inference, the model produces a free-grain prediction by selecting the deepest label that can be predicted with high confidence.

The key challenge is to learn semantic and visual relations from training data with mixed labels, where supervision is incomplete and inconsistent across levels, introducing a form of weak supervision. The model must learn from partial supervision at different depths to make consistent predictions across the taxonomy. This differs from conventional settings, where labels are provided at a single level—either fine-grained (fully supervised) or coarse (weakly supervised), without modeling cross-level interactions. (Fig. 1,1-2).

To support our free-grain setting, we construct new benchmark datasets by systematically adapting existing hierarchical datasets (CUB [43], Aircraft [22], iNat21-mini [39]) to exhibit mixed-granularity supervision. To improve diversity beyond the small-scale (CUB, Aircraft) and domain-specific (iNat21-mini) datasets, we further redesign ImageNet [32] by restructuring its irregular WordNet [10] hierarchy into a consistent three-level taxonomy for hierarchical recognition (Sec. 3).

On top of these datasets, we create two complementary variants to capture both realistic annotation difficulty and varying levels of label availability: **1) Foundation-based variants** (ImageNet-F, iNat21-mini-F, CUB-F), where label depth is determined by whether large visual foundation models [28, 35] correctly predict each level. While not a perfect proxy for human annotation, these models provide a reasonable approximation: we observe that incorrect predictions at deeper levels often correspond to visually ambiguous or annotation-challenging cases, yielding realistic mixed-granularity patterns. For example, in the last column (Fig. 2), distant birds are labeled as *Bird*, mid-range ones as *Bird of prey*, and close-ups as *Bald eagle*. In addition, this naturally induces long-tailed label availability, with fine-grained labels removed more often for some classes than

others. Fig. 2 illustrates this on our ImageNet-F dataset, which exhibits a long-tailed distribution of fine-grained labels and realistic mixed granularity aligned with visual ambiguity. **2) Randomized variants** (CUB-Rand, Aircraft-Rand), where label depths are randomly assigned at varying proportions, enabling systematic evaluation under different levels of label availability. Together, these variants span a wide range of mixed-granularity scenarios, providing a comprehensive benchmark for free-grain learning.

Our setting poses a significant challenge for existing hierarchical classifiers. When trained under free-grain supervision, state-of-the-art (SOTA) methods [7, 26] drop by up to **40%** in full-path accuracy on iNat21-mini [39], where a prediction is counted as correct only if all levels of the taxonomy are correctly predicted. This highlights the difficulty of learning from mixed-granularity labels and the need for more robust approaches.

To address this, we propose two simple yet effective methods that approach the problem from different perspectives: one leveraging semantics, the other visual structure. **1) Text-guided pseudo attributes** use a vision–language model [9] to generate image descriptions, providing semantic cues that help learn discriminative features shared across images without labels. **2) Taxonomy-guided semi-supervised learning** treats missing levels as unlabeled and exploits hierarchical consistency to learn from both labeled and unlabeled data. Across datasets, each method improves over SOTA hierarchical methods by 5–25%, providing strong baselines and highlighting substantial room for future advances in free-grain learning.

Lastly, we study **free-grained inference**, where the model adaptively selects prediction depth, motivated by the fact that a correct coarse prediction can be preferable to an incorrect fine-grained one. We consider two strategies: 1) confidence-based, selecting the deepest label with sufficient confidence, and 2) consistency-based, selecting the deepest level that maintains hierarchical consistency.

Contributions. 1) We introduce free-grained hierarchical recognition, a new task with mixed-granularity supervision and adaptive prediction depth. **2)** We construct various benchmark datasets, including foundation-based and randomized variants. **3)** We propose simple yet effective methods that leverage semantic and visual structure, consistently improving over prior hierarchical classifiers.

2. Related Work

Hierarchical classification predicts the full taxonomy path for each image, requiring accurate level-wise predictions while also encouraging parent–child consistency across the hierarchy [5, 7, 26, 41]. Meanwhile, some methods use the hierarchy as an auxiliary signal for flat (fine-grained) classification, for example, to regularize feature learning [49] or to reduce the severity of fine-grained mistakes [12, 17].

Table 1. **Our task is more practical and challenging.** Free-grain learning reflects real-world annotation, where each image may have fine (F) or coarse (C) labels, and models must predict a taxonomy-consistent hierarchy. It jointly introduces class imbalance (Cls. Imb.) and level imbalance (Lvl. Imb.), along with weak and partial supervision—factors mostly studied in isolation. Evaluation considers both accuracy (Acc.) and consistency (Con.).

	Input		Output		Labels	Imbalance		Metrics	
Tasks	F.	C.	F.	C.	Avail.	Cls.	Lvl.	Acc.	Con.
Long-tailed recog.	✓	✗	✓	✗	All	✓	✗	✓	✗
Semi-supervised	✓	✗	✓	✗	*Partial*	✗	✗	✓	✗
Weakly-supervised	✗	✓	✓	✗	All	✗	✗	✓	✗
Hierarchical recog.	✓	✓	✓	✓	All	✗	✗	✓	✓
Free-grained recog.	✓	✓	✓	✓	*Partial*	✓	✓	✓	✓

Importantly, all these approaches *assume complete hierarchical supervision* is available for every training sample.

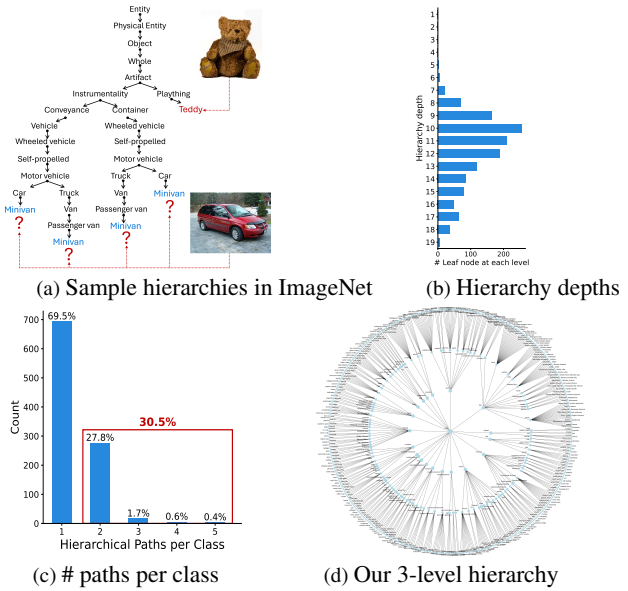
Long-tailed and semi-/weakly-supervised recognition address distinct real-world challenges [20, 25, 30, 44], but they typically operate at a *single* granularity, using either fine-grained labels alone or coarse labels alone. In contrast, our free-grain learning setting introduces a new, unexplored problem: learning from *mixed-granularity* labels within a hierarchy. This naturally brings together challenges from multiple areas, including class imbalance within each level, imbalance across different levels of the hierarchy, weak/semi-supervision, and the need to maintain hierarchical consistency, all within a single unified framework. Unlike prior work that addresses these challenges in isolation, our setting requires handling them jointly under mixed-granularity supervision. See a task comparison in Table 1. Detailed related works are provided in Appendix I.

3. Benchmarks for Free-Grained Recognition

We adapt existing hierarchical benchmarks to our free-grain setting, but prior datasets are often small-scale (e.g., CUB, Aircraft) or domain-specific (e.g., iNat21-mini), as shown in Table 2. To enable a large-scale and diverse benchmark, we reorganize ImageNet into a clean three-level hierarchy (ImageNet-3L), as its original WordNet taxonomy is irregular (Fig. 3). This simplified structure supports mixed-granularity prediction without unnecessary complexity. We first describe this restructuring (Sec. 3.1), then introduce foundation-based variants that mimic real-world annotation patterns (Sec. 3.2) and randomized variants for controlled evaluation (Sec. 3.3).

Table 2. **We convert existing hierarchical benchmarks into free-grain versions.** Since ImageNet's taxonomy is inconsistent, we newly curate a consistent three-level hierarchy, ImageNet-3L.

Dataset	**#levels**	**#classes per level**	**#train**	**#test**
CUB	3	13-38-200	5,994	5,794
Aircraft	3	30-70-100	6,667	3,333
iNat21-mini	8	3-11-13-51-273-1103-4884-10000	500,000	100,000
ImageNet	5-19	- 1000	1,281,167	50,000
ImageNet-3L	**3**	**20-127-505**	**645,480**	**25,250**



(a) Sample hierarchies in ImageNet (b) Hierarchy depths

(c) # paths per class (d) Our 3-level hierarchy

Figure 3. **We curate ImageNet-3L as a benchmark for hierarchical classification. (a)** Sample hierarchies reveal two issues: 1) some classes have multiple valid paths (e.g., *Minivan*), and 2) classes at the same depth can have mismatched specificity (e.g., *Teddy* vs. *Conveyance*). **(b)** ImageNet classes span widely varying depths (5–19 levels), often exceeding 10, highlighting inconsistency in hierarchy depth. **(c)** 30% of classes have multiple valid hierarchical paths, introducing ambiguity in evaluation. **(d)** We construct a coherent 3-level taxonomy, inspired by cognitive psychology [31]: *basic* for general recognition, *subordinate* for contextual specificity, and *fine-grained* for specialized distinctions.

3.1. Constructing ImageNet-3L

As in Fig. 3, the WordNet hierarchy [10] is noisy and inconsistent, making ImageNet unsuitable for full-path evaluation. To address this, we simplify the WordNet hierarchy by removing overly abstract nodes (e.g., *Entity*, *Whole*) and restructuring it into a consistent three-level taxonomy. This design is guided by categorization principles [31], where the *basic* level is the most natural and visually distinctive.

We anchor categories at a shared *basic* level (e.g., *vehicle, dog*) and organize subordinate and fine-grained categories under it. When multiple candidates exist, we select those that best match the granularity of existing basic classes. We further apply additional design principles, described below, with full details provided in the appendix A.

1) *Enforce meaningful structure:* We remove paths where each node has only one child, since coarse labels fully determine the fine labels. Branches with fewer than three levels are also excluded. **2) *Maximize within-group diversity:*** Among subordinate candidates under each basic class, we favor those with richer fine-grained subclasses, for example choosing *parrot* (4 children) over *cockatoo* (1 child). **3) *Refine vague categories:*** Ambiguous groups such as *Women's Clothing* are reorganized into precise, functionally grounded categories (e.g., *Underwear*) to improve clarity. **4) *Validate with language models and human review:*** We use large language models (ChatGPT [1]) to suggest refinements, with all decisions manually reviewed for semantic consistency. Applying this curation process to ImageNet-1k yields a structured benchmark of 20 basic, 127 subordinate, and 505 fine-grained classes, ImageNet-3L, ensuring every branch supports meaningful hierarchical prediction (a complete list is provided in Appendix B).

3.2. Foundation-based Pruning

To build realistic free-grain training sets, we prune hierarchical labels using large vision–language models: CLIP [28] for ImageNet-F and BioCLIP [35] for iNat21-mini-F and CUB-F. While these models are not designed to measure ambiguity, their zero-shot confidences indicate when fine-grained labels are less reliable, whether due to limited visual detail, annotation difficulty, or inconsistent expertise. Fig. 2 shows that this results in mixed-granularity supervision patterns that often follow such visual ambiguity, with distant or less discernible instances labeled more coarsely and clearer ones labeled more finely.

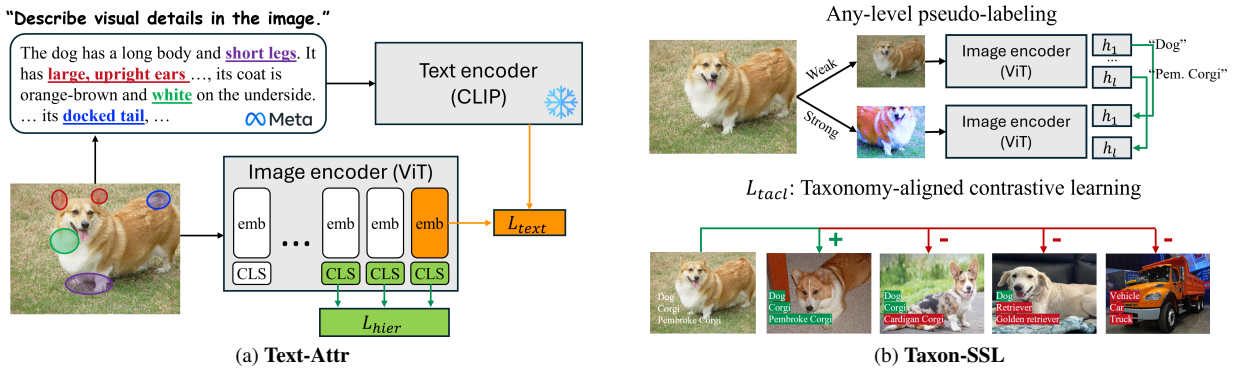
We adopt CLIP's prompt-ensemble strategy (e.g., *a photo of a [class]*) and assign labels from coarse to fine based on prediction correctness: **(1)** We always retain the *basic* label. **(2)** If the subordinate prediction is correct, we retain the *subordinate* label. **(3)** If both subordinate and fine-grained predictions are correct, we retain the *fine-grained* label. This defines the deepest available label for each image, with higher-level labels assumed from the given taxonomy. Since relying solely on foundation models' predictions can produce biased label distributions, we further remove a portion of subordinate labels based on the fine-grained removal rate per class, introducing more challenging supervision.

This pruning only removes labels and introduces no additional semantic information; it may even increase difficulty by discarding labels of harder examples. While we use CLIP and BioCLIP, similar pruning can be performed with other foundation models or ensembles (e.g., removing labels consistently mispredicted across models).

1) ImageNet-F. After pruning, 32.6% of images retain all three levels (Basic + Subordinate + Fine-grained), 28.0% retain two (Basic + Subordinate), and 39.4% retain only the Basic. Each class keeps the same number of images as in ImageNet; imbalance arises only from label granularity.

2) iNat21-mini-F. BioCLIP, a biology foundation model, performs well on species-level prediction but struggles with coarser labels. This mismatch enables substantial pruning: 22.5% of images retain all three levels (Order + Family + Species), 28.0% retain two, and 49.5% retain only Order.

3) CUB-F. With the same procedure, 31.5% of images keep 3 levels, 23.3% two (Order, Family), 45.2% only Order.



(a) **Text-Attr** (b) **Taxon-SSL**

Figure 4. **Overview of the proposed methods.** (a) Text-Attr enriches feature representations using semantic cues from images, compensating for missing labels and capturing shared attributes across levels. (b) Taxon-SSL handles missing-level labels by treating them as unlabeled and learns from visual consistency through augmented views. Both methods offer distinct benefits for our challenging task.

3.3. Randomized Pruning

To control label availability, we construct randomized variants, CUB-Rand and Aircraft-Rand, by randomly pruning labels from CUB [43] and Aircraft [22]. Unlike realistic pruning, this design systematically varies supervision and simulates *extreme* sparsity (e.g., only 10% fine-grained labels), enabling stress-testing of model robustness across diverse label distributions. Although random removal is independent of image difficulty, it reflects practical factors such as annotator expertise, cost, or task-specific constraints. We denote availability as a-b-c, where a% of basic, b% of subordinate, and c% of fine-grained labels are retained (e.g., 100-50-10 retains 10% fine-grained labels and 40% subordinate-only labels).

4. Free-Grain Learning Methods

We first define the problem (Sec. 4.1) and then present two approaches: (1) semantic guidance via text-guided pseudo attributes to learn shared visual features (Sec. 4.1), and (2) taxonomy-guided semi-supervised learning (Taxon-SSL) to leverage missing labels (Sec. 4.3).

4.1. Problem Setup

In free-grained hierarchical classification, the goal is to predict labels across all levels of a taxonomy from training data with mixed granularity. Each image is annotated at a certain level, and all coarser labels are assumed to be available while finer ones are missing; the coarsest label is always given. The model is trained to produce consistent predictions across the full hierarchy.

Free-grained Hierarchical Loss. To adapt prior hierarchical recognition methods to the free-grain setting, we modify their hierarchical supervision by applying the loss only at levels with available labels. Given hierarchical labels $y_1, \ldots, y_L$ across L levels, the loss is defined as:

$$\mathcal{L}_{\text{hier}} = \sum_{l=1}^{L} \mathbb{1}_{\{y_l \text{ exists}\}} \cdot \mathcal{L}(f_l(x), y_l), \tag{1}$$

where $f_l(x)$ is the prediction at level l, and $\mathcal{L}$ denotes a classification loss (e.g., cross-entropy).

4.2. Semantic Guidance: Text-Guided Pseudo Attributes (Text-Attr)

Our semantic guidance approach is motivated by the observation that while class labels differ across hierarchical levels (e.g., *Dog* $\rightarrow$ *Corgi* $\rightarrow$ *Pembroke*), many visual attributes, such as "tail length" or "ear shape", remain consistent (Fig. 4a). To capture these shared semantic cues, we use image descriptions as auxiliary supervision. While recent large language models (LLM) (e.g., ChatGPT [1])-based approaches such as FineR [19] also use vision-language model (VLM)-generated text, their purpose is different: they feed these cues into an LLM for training-free fine-grained class reasoning, whereas we use text as supervision to train image representations to capture visual attributes shared across hierarchical levels.

Specifically, given an input image x, we use a frozen vision-language model (VLM), Llama-3.2-11B [9], to generate a language description d_x, using the prompt: "*Describe visual details in the image.*" This produces descriptions containing phrases such as "short legs" or "pointed ears," which we encode into a text embedding z_x^t using CLIP's text encoder [28]. We cap generation at 100 tokens, while CLIP accepts 77 tokens; longer descriptions are truncated during encoding. Although truncation discards some details, our method focuses on shared semantic cues (e.g., "short legs," "brown markings") rather than exhaustive captions, making it robust to this limitation. In parallel, we obtain the image embedding z_x^v from the image encoder, and align it with the text embedding z_x^t using a contrastive loss:

$$\mathcal{L}_{\text{text}} = -\frac{1}{N} \sum_{i=1}^{N} \log \left(\frac{\exp(\text{sim}(z_i^v, z_i^t)/\tau)}{\sum_{j=1}^{N} \exp(\text{sim}(z_i^v, z_j^t)/\tau)} \right), \tag{2}$$

where $\text{sim}(\cdot, \cdot)$ is cosine similarity and τ is a temperature

parameter. This loss guides the encoder to capture salient, label-independent traits shared across levels. Although not explicitly predicting attributes, aligning image features with text induces intermediate representations, which we call pseudo-attributes. This model-agnostic method can be applied to any architecture.

4.3. Visual Guidance: Taxonomy-guided Semi-Supervised Learning (Taxon-SSL)

We adopt a semi-supervised formulation because missing-grain labels can be treated as unlabeled data. CHMatch [44] shows that coarse labels can improve pseudo-labeling, but it is limited to a two-level (coarse–fine) setting and focuses on refining fine-grained predictions. We generalize this to arbitrary multi-level taxonomies by 1) generating pseudo-labels at *every level*, and 2) enforcing *cross-level consistency* so that predictions remain valid along the hierarchy.

1) Multi-level pseudo-labeling. Following CHMatch, we decouple the classifier f into a shared feature extractor f_{feat} and level-specific heads $\{h_l\}_{l\in\mathcal{S}_x}$, where each head predicts labels at a different taxonomy level. The supervised loss is computed using Eq. 1, applying supervision only at levels with available labels. Pseudo-labels at each level are generated from the predictions of the corresponding head given a weakly augmented input $W(x)$.

2) Taxonomy-aligned feature learning. A key challenge is that pseudo-labels at different levels may be inconsistent (e.g., two samples share a coarse label but differ at fine levels). To address this, we only treat pairs as *reliable positives* when they agree across *all levels*.

For each mini-batch, we build level-wise affinity graphs W^l based on pseudo-label agreement: $W^l_{ij} = 1$ if images i and j share the same pseudo-label at level l, and 0 otherwise. We then define a taxonomy-aligned affinity:

$$W_{ij} = \begin{cases} 1 & \text{if } W^1_{ij} = \cdots = W^L_{ij} = 1, \\ 0 & \text{otherwise.} \end{cases} \tag{3}$$

This enforces that two samples are considered similar *only if they are consistent across the entire hierarchy*, effectively filtering out noisy or partially incorrect pseudo-labels.

Contrastive objective. Using W, we pull together positive pairs ($W_{ij} = 1$) and push apart negative pairs ($W_{ij} = 0$). The taxonomy-aligned contrastive loss is:

$$\mathcal{L}_{\text{tacl}} = -\frac{1}{\sum_j W_{ij}} \log \frac{\sum_j W_{ij} \exp\left((g(f(x_i)) \cdot g(f(x_j))')/t\right)}{\sum_j (1 - W_{ij}) \exp\left((g(f(x_i)) \cdot g(f(x_j))')/t\right)}, \tag{4}$$

where $g(f(x_i))$ is the projected feature of image i, and t is a temperature. This objective encourages samples with consistent hierarchical semantics to form tight clusters in the feature space, while separating those that disagree at any level. See more details in the Appendix H.

5. Experiments

In this section, we first describe the experimental setup (Sec. 5.1). We then present results on our free-grain benchmarks (Sec. 5.2) and provide further analysis of the proposed methods (Sec. 5.3). Finally, we compare free-grained inference methods (Sec. 5.4).

5.1. Experimental Setup

1) Dataset: We conduct experiments using our proposed *ImageNet-F*, *iNat21-mini-F*, and *CUB-F* datasets, along with the synthetic *CUB-Rand* and *Aircraft-Rand* datasets.

2) Evaluation metrics: Following [26], we evaluate accuracy and consistency: *(1) Level-accuracy*: Top-1 accuracy at each level. *(2) Tree-based InConsistency Error rate (TICE)*: Proportion of samples with inconsistent predictions in the hierarchy (lower is better): TICE $= \frac{n_{\text{ic}}}{N}$, where N is the total number of samples and n_{ic} is the number of inconsistent predictions. *(3) Full-Path Accuracy (FPA)*: Proportion of samples correctly predicted at all levels (primary metric): FPA $= \frac{n_{\text{ac}}}{N}$, where n_{ac} is the number of samples correct at all hierarchy levels.

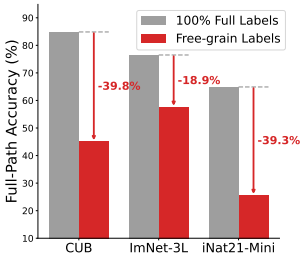
3) Comparison Methods: We adapt two strong and relevant hierarchical classifiers to the free-grain setting for comparison. *(1) Hierarchical Residual Network (HRN)* [7]: the first to handle supervision at both subordinate and fine-grained levels by maximizing marginal probabilities within the tree-constrained space. *(2) H-CAST* [26]: the current SOTA, encouraging consistent visual grouping across taxonomy levels. Originally trained with full supervision, we adapt it to this setting via the level-wise loss in Eq. 1, using only available labels.

4) Implementation: We use H-ViT, a ViT-Small-based hierarchical classifier, as the backbone for evaluating both Text-Attr and Taxon-SSL. To evaluate its compatibility across architectures, we also apply Text-Attr to H-CAST [26], a state-of-the-art hierarchical model with comparable capacity. HRN [7] is evaluated with its original ResNet-50 backbone, which has over twice the parameters. All models are trained for 100 epochs, except for ImageNet-F, which is trained for 200 due to its larger scale. Full architectural and training details are in the appendix G.

5.2. Benchmarking Results

Result 1: Performance drop under free-grain learning. The prior hierarchical SOTA, H-CAST, degrades sharply under mixed-granularity labels on both CUB and iNat21-mini. Fig. 5a shows that the full-path accuracy drops from 84.9% to 45.1% on CUB-F and from 64.9% to 25.6% on iNat21-mini-F. This highlights the challenge of mixed-granularity labels and imbalanced supervision across the hierarchy, and the need for methods that handle them.

Result 2: Performance on ImageNet-F. As shown in Ta-



(a) **H-CAST: full vs. free-grain**

Dataset	ImageNet-F (20-127-505)					iNat21-mini-F (273 - 1,103 - 10,000)				
	FPA(↑)	fine.(↑)	sub.(↑)	basic(↑)	TICE(↓)	FPA(↑)	spec.(↑)	fam.(↑)	order(↑)	TICE(↓)
HRN [7]	37.79	38.73	55.73	78.65	46.69	17.03	25.43	46.51	70.20	53.81
H-CAST [26]	57.59	59.02	82.69	93.53	21.81	25.63	28.61	67.20	83.62	47.17
Taxon-SSL	48.40	52.34	65.74	82.96	19.87	31.74	**37.11**	69.53	82.02	37.31
Taxon-SSL + Text-Attr	49.65	53.43	66.43	83.56	18.81	**31.93**	37.08	69.76	82.20	**37.04**
Text-Attr (H-ViT)	55.48	59.05	77.95	89.45	24.02	27.88	32.07	68.27	80.49	46.35
Text-Attr (H-CAST)	**63.20**	**64.91**	**84.47**	**93.56**	**18.58**	29.74	32.37	**71.79**	**85.99**	44.63

(b) **Method comparison on ImageNet-F and iNat21-mini-F.**

Figure 5. **(a) Transitioning from fully labeled data to our free-grain setting results in a substantial drop in Full-Path Accuracy, highlighting the difficulty of the task.** SOTA H-CAST drops by 19–40 percentage points across datasets. **(b) Our methods effectively improve performance under free-grain supervision, with behavior depending on data characteristics.** Conventional hierarchical methods such as HRN [7] and H-CAST [26] degrade significantly under incomplete supervision. In contrast, Text-Attr (H-CAST) performs strongly on ImageNet-F, where rich visual cues support text-guided learning, while Taxon-SSL is more effective on iNat21-mini-F, where fine-grained classes have similar appearances. Combining both (Taxon-SSL + Text-Attr) yields consistent but modest gains across datasets.

ble 5b, existing hierarchical methods degrade sharply under free-grain learning: HRN reaches only 37.8% FPA, while H-CAST performs better at 57.6% but still struggles with missing labels. Text-Attr (H-ViT) achieves 55.5% without relying on H-CAST's visual grouping, and integrating it into H-CAST further improves performance to 63.2%, demonstrating the effectiveness of semantic-guided pseudo-attribute learning at scale. Taxon-SSL improves over HRN by leveraging visual guidance but remains less effective than Text-Attr methods, whose strong performance benefits from the abundance and diversity of ImageNet-F for reliable visual–semantic alignment.

Result 3: Performance on iNat21-mini-F. In Table 5b, on the large-scale iNat21-mini-F dataset, which contains many classes (10,000), conventional hierarchical methods perform poorly (17.0% for HRN, 25.63% for H-CAST). Taxon-SSL achieves the best performance (31.9% FPA), highlighting the benefits of structural label propagation under limited per-class supervision. Text-Attr methods perform slightly lower (27.9–30.0% FPA), likely due to restricted textual diversity in this fine-grained biological domain, yet still outperform conventional baselines.

Additional results and ablations. We report additional results on CUB-F (Sec. D.1) and randomized variants with varying (limited) label availability (Sec. D.2). Across these settings, conventional hierarchical methods degrade under mixed-granularity supervision, while our approaches remain effective and robust. We further evaluate robustness under the original (unrefined) WordNet hierarchy, which exhibits irregular depth and inconsistent granularity (Sec. C). In addition, we conduct ablations on text encoders, Text-Attr features, training strategies, and architecture design (Sec. F), validating each component's contribution.

5.3. Further Analysis

How do methods behave with varying label availability? Text-attr excels with sparse labels, Taxon-SSL with moderate label availability. We analyze class-wise performance under imbalanced fine-grained label availability on ImageNet-F. To isolate effects, we compare Text-Attr (H-ViT) and Taxon-SSL with identical ViT-small backbones, excluding H-CAST modules. Fig. 6 plots per-class accuracy, where the x-axis shows Text-Attr performance and the y-axis shows Taxon-SSL performance; the diagonal indicates equal performance. Text-Attr (H-ViT) tends to outperform in label-scarce classes, appearing below the diagonal, by leveraging textual descriptions as additional supervision, while Taxon-SSL performs better for classes with more training samples, appearing above the diagonal by propagating consistency across missing levels. We provide additional t-SNE [21] analysis in Appendix E.

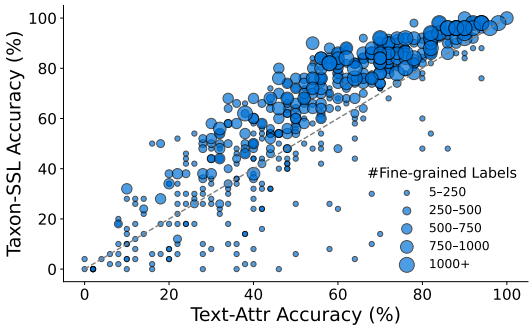


Figure 6. **Text-Attr is more effective under scarce fine-grained supervision, while Taxon-SSL performs better with more training data.** Each circle represents a class in ImageNet-F, with Text-Attr fine-grained accuracy on the x-axis and Taxon-SSL accuracy on the y-axis. The diagonal marks equal performance: points below favor Text-Attr, and points above favor Taxon-SSL. Circle size indicates the number of available training samples per class. Smaller circles tend to lie below the diagonal, showing the advantage of Text-Attr under limited data by leveraging textual guidance, whereas larger circles more often lie above it, indicating that Taxon-SSL benefits from richer supervision.

How does external semantic guidance help? External semantic guidance helps the model attend to semantically relevant features and improves hierarchical consistency. To as-

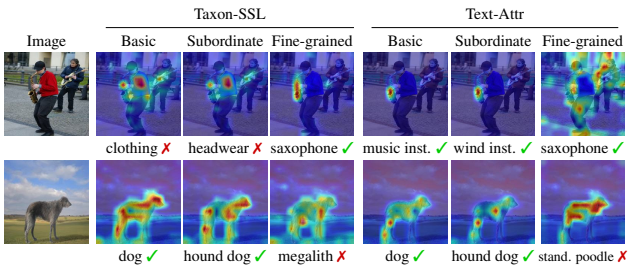


Figure 7. **Text-Attr improves semantic focus under diverse large-scale data.** (**Row 1**) In a multi-object image, Taxon-SSL assigns inconsistent labels ("*clothing*" at the basic level, "*saxophone*" at the fine-grained level), while Text-Attr (H-ViT) correctly predicts "*musical instrument*" by focusing on the relevant object. (**Row 2**) When both fail at the fine-grained level, Taxon-SSL outputs an unrelated class ("*megalith*"), whereas Text-Attr (H-ViT) chooses a semantically closer one ("*poodle*"). This shows that text-derived attributes help the model attend to meaningful regions and maintain semantic plausibility, on a large-scale ImageNet-F dataset with diverse categories and sparse labels. Green/Red denote correct/incorrect predictions.

sess this effect, we compare saliency maps [6] from Taxon-SSL and Text-Attr (H-ViT) (Fig. 7). In Row 1, with multiple objects, Taxon-SSL focuses on a human shoulder and misclassifies the image, violating the hierarchy, while Text-Attr attends to the instrument and predicts correctly. In Row 2, when both fail at the fine-grained level, Taxon-SSL predicts an unrelated class, whereas Text-Attr selects a visually similar dog by focusing on curly fur and body shape. These results show that text-derived semantic cues guide attention toward meaningful features across label granularities, while Taxon-SSL may drift to visually salient but semantically irrelevant regions under sparse or ambiguous supervision.

5.4. Free-grained Inference

Free-grained inference is important in practice, as a correct coarse label is often preferable to an incorrect fine-grained one. We compare two stopping strategies: confidence-based and consistency-based rules. Confidence-based stopping uses softmax probability with a threshold $\tau = 0.9$ (selected from $[0.85, 0.99]$), halting when $P(y|x) < \tau$. However, as shown in Fig. 8, it often stops prematurely because probability is spread across similar sibling classes. In contrast, consistency-based stopping halts only when taxonomy constraints are violated, requiring no threshold tuning and more reliably reaching deeper correct levels. As a result, it produces more reliable and deeper correct predictions than confidence-based rules. Under consistency-based stopping, Text-Attr (H-CAST) produces the most taxonomy-consistent predictions (Fig. 9), reaching deeper correct levels while avoiding inconsistent fine-grained outputs (e.g., stopping at "*dog* → *hound*" when the fine label is incorrect). This shows that stronger hierarchical consistency leads to more effective free-grained inference.

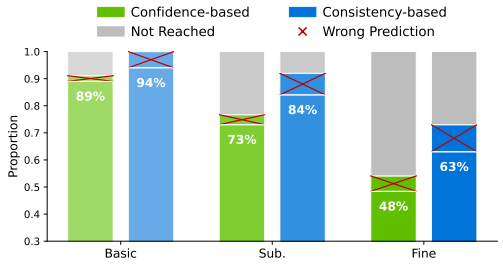
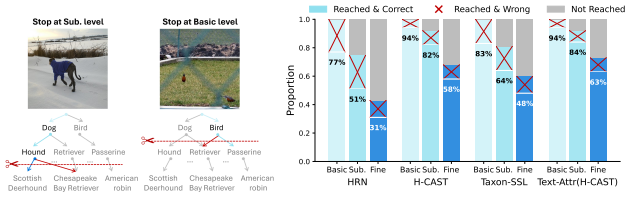


Figure 8. **Confidence vs. consistency stopping (ImageNet-F, Text-Attr (H-CAST)).** Consistency-based stopping (blue) yields more reliable and deeper correct predictions than confidence-based stopping (green). Bars show the proportion of samples reaching each level (Basic → Subordinate → Fine). Gray ("Not Reached") indicates early stopping at a coarser level, and red crosses mark incorrect predictions. Confidence-based rules often stop early, failing to reach deeper levels due to probability splitting among similar classes, whereas consistency-based stopping more often reaches deeper correct predictions.



(a) Examples: Text-Attr (H-CAST) (b) Methods under consistency-based stopping

Figure 9. **Free-grained inference results with consistency-based stopping on ImageNet-F. (a)** The model stops at the appropriate level-subordinate (e.g., *Hound*, left) or basic (e.g., *Bird*, right)—when deeper predictions become inconsistent, yielding more reliable outputs. **(b)** Inference stops whenever finer-level predictions conflict with their coarser ancestors. On ImageNet-F, Text-Attr (H-CAST) descends deeper into the hierarchy while maintaining correctness, whereas HRN halts earlier and produces fewer fine-level predictions.

6. Conclusion

We introduce free-grained hierarchical recognition, where models learn from labels of varying granularity while maintaining taxonomy consistency. To support this setting, we present diverse benchmarks and two effective baselines. While our methods achieve strong performance, there remains significant room for improvement, highlighting the challenges of free-grain learning and motivating the development of more robust approaches.

Limitations: Our methods do not explicitly model class- or level-wise imbalance; future work could explore imbalance-aware strategies for further improvement. Additionally, CLIP-based pruning is not a perfect proxy for visual ambiguity, but serves as a practical approximation of label difficulty, with ensemble-based approaches potentially improving the pruning process.

Acknowledgements

This project was supported, in part, by NSF 2215542, NSF 2313151, and Bosch gift funds to S. Yu at UC Berkeley and the University of Michigan, with additional compute support from NAIRR Pilot (CIS250430, CIS240431).

References

[1] Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023. 4, 5, 12, 28

[2] Sumyeong Ahn, Jongwoo Ko, and Se-Young Yun. CUDA: Curriculum of data augmentation for long-tailed recognition. In *The Eleventh International Conference on Learning Representations*, 2023. 28

[3] Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, et al. Qwen2. 5-vl technical report. *arXiv preprint arXiv:2502.13923*, 2025. 28

[4] David Berthelot, Nicholas Carlini, Ian Goodfellow, Nicolas Papernot, Avital Oliver, and Colin A Raffel. Mixmatch: A holistic approach to semi-supervised learning. *Advances in neural information processing systems*, 2019. 28

[5] Dongliang Chang, Kaiyue Pang, Yixiao Zheng, Zhanyu Ma, Yi-Zhe Song, and Jun Guo. Your" flamingo" is my" bird": fine-grained, or not. In *CVPR*, 2021. 1, 3, 18, 28

[6] Hila Chefer, Shir Gur, and Lior Wolf. Transformer interpretability beyond attention visualization. In *CVPR*, 2021. 8

[7] Jingzhou Chen, Peng Wang, Jian Liu, and Yuntao Qian. Label relation graphs enhanced hierarchical residual network for hierarchical multi-granularity classification. In *CVPR*, 2022. 1, 3, 6, 7, 18, 20, 21, 26, 28

[8] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020. 26

[9] Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, et al. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*, 2024. 3, 5

[10] Christiane Fellbaum. *WordNet: An electronic lexical database*. MIT press, 1998. 2, 4, 28

[11] Ashima Garg, Shaurya Bagga, Yashvardhan Singh, and Saket Anand. Hiermatch: Leveraging label hierarchies for improving semi-supervised learning. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 2022. 28

[12] Ashima Garg, Depanshu Sani, and Saket Anand. Learning hierarchy aware features for reducing mistake severity. In *European Conference on Computer Vision*, 2022. 3, 28

[13] Matej Grcic, Artyom Gadetsky, and Maria Brbic. Fine-grained classes and how to find them. In *Proceedings of the 41st International Conference on Machine Learning*, 2024. 1, 28

[14] Taegil Ha, Seulki Park, and Jin Young Choi. Novel regularization via logit weight repulsion for long-tailed classification. In *34th British Machine Vision Conference 2023, BMVC 2023, Aberdeen, UK, November 20-24, 2023*, 2023. 28

[15] Aaron Hurst, Adam Lerer, Adam P Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, et al. Gpt-4o system card. *arXiv preprint arXiv:2410.21276*, 2024. 28

[16] Juan Jiang, Jingmin Yang, Wenjie Zhang, and Hongbin Zhang. Hierarchical multi-granularity classification based on bidirectional knowledge transfer. *Multimedia Systems*, 2024. 1, 28

[17] Shyamgopal Karthik, Ameya Prabhu, Puneet K. Dokania, and Vineet Gandhi. No cost likelihood manipulation at test time for making better mistakes in deep networks. In *International Conference on Learning Representations*, 2021. 3, 28

[18] Dong-Jin Kim, Zhongqi Miao, Yunhui Guo, and X Yu Stella. Modeling semantic correlation and hierarchy for real-world wildlife recognition. *IEEE Signal Processing Letters*, 2023. 1, 28

[19] Mingxuan Liu, Subhankar Roy, Wenjing Li, Zhun Zhong, Nicu Sebe, and Elisa Ricci. Democratizing fine-grained visual recognition with large language models. In *The Twelfth International Conference on Learning Representations*, 2024. 5, 28

[20] Ziwei Liu, Zhongqi Miao, Xiaohang Zhan, Jiayun Wang, Boqing Gong, and Stella X. Yu. Large-scale long-tailed recognition in an open world. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019. 3, 28

[21] Laurens van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. *Journal of machine learning research*, 2008. 7, 22

[22] S. Maji, J. Kannala, E. Rahtu, M. Blaschko, and A. Vedaldi. Fine-grained visual classification of aircraft. Technical report, 2013. 2, 5, 28

[23] Zhongqi Miao, Stella X Yu, Kyle L Landolt, Mark D Koneff, Timothy P White, Luke J Fara, Enrika J Hlavacek, Bradley A Pickens, Travis J Harrison, and Wayne M Getz. Challenges and solutions for automated avian recognition in aerial imagery. *Remote Sensing in Ecology and Conservation*, 2023. 1

[24] Seulki Park, Jongin Lim, Younghan Jeon, and Jin Young Choi. Influence-balanced loss for imbalanced visual classification. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2021. 28

[25] Seulki Park, Youngkyu Hong, Byeongho Heo, Sangdoo Yun, and Jin Young Choi. The majority can help the minority: Context-rich minority oversampling for long-tailed classification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022. 3, 28

[26] Seulki Park, Youren Zhang, Stella X. Yu, Sara Beery, and Jonathan Huang. Visually consistent hierarchical image classification. In *The Thirteenth International Conference on Learning Representations*, 2025. 1, 3, 6, 7, 18, 20, 21, 26, 28

[27] Sarah Pratt, Ian Covert, Rosanne Liu, and Ali Farhadi. What does a platypus look like? generating customized prompts for zero-shot image classification. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023. 28

[28] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*. PMLR, 2021. 2, 4, 5, 28

[29] Jiawei Ren, Cunjun Yu, shunan sheng, Xiao Ma, Haiyu Zhao, Shuai Yi, and hongsheng Li. Balanced meta-softmax for long-tailed visual recognition. In *Advances in Neural Information Processing Systems*, 2020. 28

[30] Joshua Robinson, Stefanie Jegelka, and Suvrit Sra. Strength from weakness: Fast learning using weak supervision. In *Proceedings of the 37th International Conference on Machine Learning*, 2020. 3, 28

[31] Eleanor Rosch, Carolyn B Mervis, Wayne D Gray, David M Johnson, and Penny Boyes-Braem. Basic objects in natural categories. *Cognitive psychology*, 1976. 4, 12, 28

[32] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, Alexander C. Berg, and Li Fei-Fei. ImageNet Large Scale Visual Recognition Challenge. *IJCV*, 2015. 2, 28

[33] Oindrila Saha, Grant Van Horn, and Subhransu Maji. Improved zero-shot classification by adapting vlms with text descriptions. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2024. 28

[34] Kihyuk Sohn, David Berthelot, Nicholas Carlini, Zizhao Zhang, Han Zhang, Colin A Raffel, Ekin Dogus Cubuk, Alexey Kurakin, and Chun-Liang Li. Fixmatch: Simplifying semi-supervised learning with consistency and confidence. *Advances in neural information processing systems*, 2020. 28

[35] Samuel Stevens, Jiaman Wu, Matthew J Thompson, Elizabeth G Campolongo, Chan Hee Song, David Edward Carlyn, Li Dong, Wasila M Dahdul, Charles Stewart, Tanya Berger-Wolf, Wei-Lun Chao, and Yu Su. Bioclip: A vision foundation model for the tree of life. In *CVPR*, 2024. 2, 4

[36] Yuwen Tan, Yuan Qing, and Boqing Gong. Vision llms are bad at hierarchical visual understanding, and llms are the bottleneck. *arXiv preprint arXiv:2505.24840*, 2025. 28

[37] Antti Tarvainen and Harri Valpola. Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results. In *Advances in Neural Information Processing Systems*, 2017. 28

[38] Changyao Tian, Wenhai Wang, Xizhou Zhu, Jifeng Dai, and Yu Qiao. Vl-ltr: Learning class-wise visual-linguistic representation for long-tailed visual recognition. In *European conference on computer vision*. Springer, 2022. 28

[39] Grant Van Horn, Elijah Cole, Sara Beery, Kimberly Wilber, Serge Belongie, and Oisin Mac Aodha. Benchmarking representation learning for natural world image collections. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2021. 2, 3, 28

[40] Liang Wang, Nan Yang, Xiaolong Huang, Binxing Jiao, Linjun Yang, Daxin Jiang, Rangan Majumder, and Furu Wei. Text embeddings by weakly-supervised contrastive pre-training. *arXiv preprint arXiv:2212.03533*, 2022. 24

[41] Rui Wang, Cong Zou, Weizhong Zhang, Zixuan Zhu, and Lihua Jing. Consistency-aware feature learning for hierarchical fine-grained visual classification. In *Proceedings of the 31st ACM International Conference on Multimedia*, 2023. 1, 3, 28

[42] Xudong Wang, Long Lian, Zhongqi Miao, Ziwei Liu, and Stella Yu. Long-tailed recognition by routing diverse distribution-aware experts. In *International Conference on Learning Representations*, 2021. 28

[43] P. Welinder, S. Branson, T. Mita, C. Wah, F. Schroff, S. Belongie, and P. Perona. Caltech-UCSD Birds 200. Technical report, California Institute of Technology, 2010. 2, 5, 28

[44] Jianlong Wu, Haozhe Yang, Tian Gan, Ning Ding, Feijun Jiang, and Liqiang Nie. Chmatch: contrastive hierarchical matching and robust adaptive threshold boosted semi-supervised learning. In *CVPR*, 2023. 3, 6, 26, 27, 28

[45] Tz-Ying Wu, Pedro Morgado, Pei Wang, Chih-Hui Ho, and Nuno Vasconcelos. Solving long-tailed recognition with deep realistic taxonomic classifier. In *ECCV*, 2020. 28

[46] Seokju Yun and Youngmin Ro. Shvit: Single-head vision transformer with memory efficient macro design. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5756–5767, 2024. 1

[47] Siqi Zeng, Remi Tachet des Combes, and Han Zhao. Learning structured representations by embedding class hierarchy. In *The Eleventh International Conference on Learning Representations*, 2022. 28

[48] Xiaohua Zhai, Basil Mustafa, Alexander Kolesnikov, and Lucas Beyer. Sigmoid loss for language image pre-training. In *Proceedings of the IEEE/CVF international conference on computer vision*, 2023. 24

[49] Shu Zhang, Ran Xu, Caiming Xiong, and Chetan Ramaiah. Use all the labels: A hierarchical multi-label contrastive learning framework. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022. 3, 28

[50] Qihao Zhao, Yalun Dai, Hao Li, Wei Hu, Fan Zhang, and Jun Liu. Ltgc: Long-tail recognition via leveraging llms-driven generated content. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024. 28

[51] Zhaoheng Zheng, Jingmin Wei, Xuefeng Hu, Haidong Zhu, and Ram Nevatia. Large language models are good prompt learners for low-shot image classification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024. 28