VISUALLY CONSISTENT HIERARCHICAL IMAGE CLASSIFICATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Hierarchical classification requires predicting an entire taxonomy tree rather than a single flat level, which demands both accurate predictions at each level and consistency across levels. However, solving hierarchical classification often compromises fine-grained accuracy compared to flat classification because each level requires distinct features, making it a multi-task problem. For example, the fine-grained classification of "Green Hermit" and "Ruby-throated Hummingbird" demands more specific details, while distinguishing between "bird" and "plant" at the coarse level requires broader features. Prior methods address this by using separate blocks for each level to learn distinct features. However, this approach struggles to resolve inconsistencies, as classifiers tend to focus on different, unrelated regions.

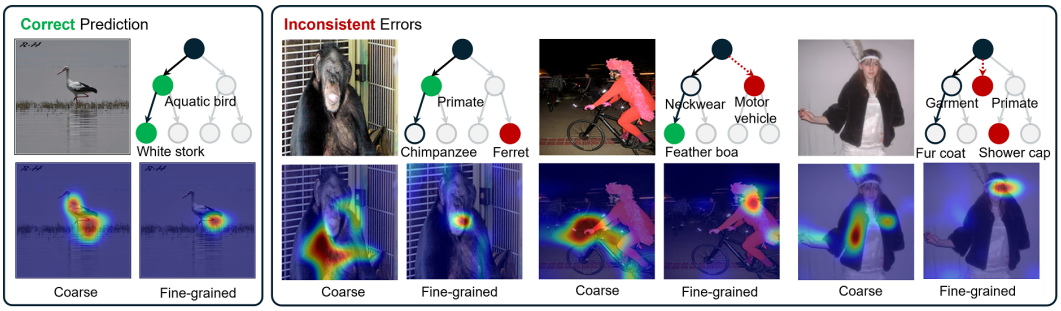
Our key insight is that classifiers across levels should be grounded in consistent visual cues. For example, the fine-grained classifier may focus on details such as the beak and wings to identify a "Green Hermit", and then the coarse classifier identifies "bird" by grouping these details into the overall "bird" shape. Therefore, we propose a novel hierarchical model that grounds fine-to-coarse *semantic parsing* on consistent hierarchical *visual segmentation*. We also introduce a tree-path KL divergence loss to enforce semantic consistency across levels. Our approach significantly outperforms zero-shot CLIP and other state-of-the-art methods on common hierarchical classification benchmarks.

1 INTRODUCTION

Hierarchical classification (Chang et al., 2021; Chen et al., 2022; Jiang et al., 2024) predicts labels at multiple levels of granularity (e.g. "Birds"-"Hummingbird"-"Green Hermit") and is crucial in real-world applications. The required level of detail varies by user expertise: non-experts may only need a coarse label like "Bird," while experts, such as biologists, require fine-grained predictions, "Green Hermit". Moreover, fine-grained flat classification often fails when details are unclear, such as when observing birds flying at high altitudes. In such cases, a model capable of multi-granularity predictions, including higher-level labels, offers greater robustness and adaptability.

Hierarchical classification presents several challenges, primarily because it requires recognizing objects at different levels of detail, turning it into a multitask problem. For example, the features necessary to distinguish between coarse categories (e.g., "Birds" vs. "Plants") may differ significantly from those needed to differentiate fine-grained categories (e.g., "Green Hermit" vs. "Ruby-throated Hummingbird"). At the coarse level, the model may need broader shapes, while at the fine-grained level, it will need finer details like the bird's beak or feather patterns. Thus, simply applying a loss to match labels at each level can lead to conflicting optimization objectives, potentially harming fine-grained performance compared to flat classification (Chang et al., 2021).

To address the reduced performance at the fine-grained level, state-of-the-art methods design separate blocks for each level of the hierarchy (Chang et al., 2021; Chen et al., 2022; Wang et al., 2023a). While these approaches achieve good results at the fine-grained level, using separate blocks for each hierarchy level makes it difficult to resolve inconsistent predictions — where predictions do not adhere to the hierarchical taxonomy. Figure 1 (b, c, d) shows the examples of inconsistent predictions from training FGN (Chang et al., 2021) on the 2-level hierarchy dataset, BREEDS (Entity-30) (Santurkar et al., 2021). For example, in the top row of image (c), the coarse classifier predicts the image



(a) Coarse O & Fine O (b) Coarse O & Fine X (c) Coarse X & Fine O (d) Coarse X & Fine X

Figure 1: **To address inconsistent predictions in hierarchical classification, classifiers at different levels should be grounded in consistent visual cues.** When coarse and fine-grained predictions are inconsistent at test time, neither can be trusted without knowing the true label, making both predictions unusable. Thus, it is crucial to develop models that maintain consistency across all levels. We present examples of predictions from training FGN (Chang et al., 2021) on BREEDS dataset (Entity-30) (Santurkar et al., 2021). In cases of inconsisteny (b, c, d), we observe through Grad-CAM visualization (Selvaraju et al., 2017) that each classifier focuses on entirely different regions (bottom). In contrast, in (a), both classifiers target the same object, with the fine-grained classifier emphasizing specific details and the coarse classifier covering a broader context. From this observation, we propose a model with consistent visual grounding, where both classifiers attend to the same target but capture different details. Our model successfully achieves correct predictions across all cases (a-d).

as a "motor vehicle", while the fine-grained classifier predicts it as a "feather boa", illustrating an inconsistent prediction.

To understand the reasons behind these inconsistencies, we use Grad-CAM visualization (Selvaraju et al., 2017) and **newly discover that the coarse and fine-grained classifiers separately attend to different areas** (Figure 1, bottom). For instance, in the case of (c), the fine-grained classifier attend to the part of the "feather boa", while the coarse classifier focuses on the "bicycle's handlebars and frame". Similarly, in Figure 1 (d), the model misclassifies the object as a "primate" by focusing on the black fur and feather in the hair at the coarse level. At the fine-grained level, it misidentifies the object as a "shower cap" by focusing on the forehead. In contrast, in the consistently correct prediction case (Figure 1 (a)), the fine-grained classifier focuses on the tail and leg areas to distinguish the "white stork" from other bird species, while the coarse classifier recognizes the overall shape of the bird, including the leg area, that the fine-grained classifier also attends to. We further quantitatively validate our observation in Appendix A.

From this observation, we propose a novel method that align the focus areas of the fine-grained and coarse classifiers, rather than training them as separate blocks. **Our insight is that, in hierarchical classification, the coarse and fine-grained classifiers should maintain visual consistency, allowing them to observe the same object *at different levels of detail*.** This concept is illustrated in Figure 2. For instance, while the fine-grained classifier examines features such as the *beak*, *wings*, and *tail* to classify "Green Hermit", the coarse classifier differentiates between "birds" and "plants" by integrating these details into the overall *body* of the bird. Therefore, we propose a new integrated model, which groups fine-grained details into increasingly coarse shapes and transfers features learned at each hierarchy level to the next. This progressive learning scheme can effectively address inconsistencies in hierarchical classification compared to previous methods learning each feature at an independent block.

To identify and group the visual details in an image, we employ the recent unsupervised image segmentation method, CAST (Ke et al., 2024). CAST has demonstrated the capability to group related pixels consistently through internal parsing within images, *without* segmentation labels. Inspired by this, we propose *Hierarchical-CAST* (H-CAST), which utilizes fine-to-coarse semantic parsing to align the focus of different levels of classifiers on the same areas for hierarchical classification. To the best of our knowledge, our work is the first to address visual consistency in hierarchical classification by utilizing unsupervised semantic segments. Since our method is an integrated model, if details initially captured at the fine-grained level are incorrect, it will receive negative signals (errors) during the learning process toward coarser levels. As training processes, the model is encouraged to capture accurate fine-grained details to improve learning at subsequent levels.

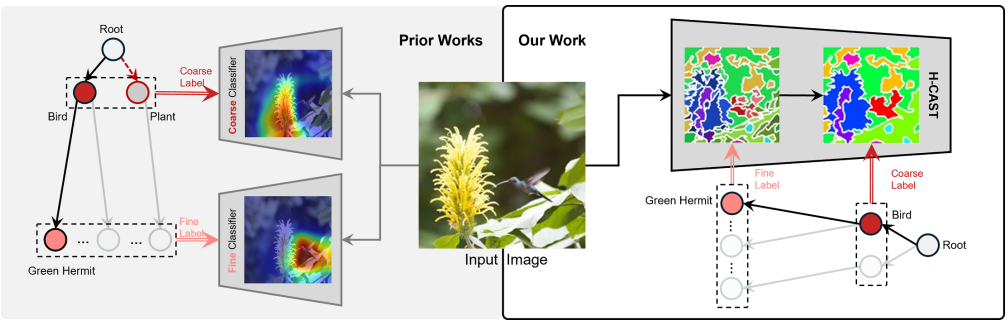


Figure 2: **While prior methods use separate blocks for different levels and struggle with inconsistencies, our integrated model ensures consistent predictions by grouping fine-grained details into coarser representations and promoting internal visual parsing.** In the segmentation images inside H-CAST, consistent groupings are represented by identical colors. The segmentation images are the results of 32-way and 8-way, respectively. For instance, when observing the red bird segments, we notice finer details such as wings, body, head, and tail in the fine level (32-way), while in the subsequent level (8-way), it is grouped as the entire bird. By leveraging this consistent internal parsing, we encourage the model to focus on the coherent regions within images.

Additionally, we propose Tree-path KL Divergence loss to further enhance semantic consistency by considering label relationships across levels so that predictions at the fine-grained level and coarse level align within the taxonomy. Our proposed method can achieve consistently correct predictions across all cases in Figure 1.

Firstly, we evaluate CLIP (Radford et al., 2021) and show that even vision foundation models struggle with inconsistent hierarchical classification. Then, to assess both accuracy and consistency, we evaluate our method with a new metric called Full-Path Accuracy (FPA), which measures the proportion of samples in the dataset correctly predicted at all hierarchy levels. Validated on common benchmarks for hierarchical recognition, our method consistently outperforms the state-of-the-art methods with fewer parameters and training time. We empirically demonstrate its effectiveness through extensive experiments and analysis. Furthermore, compared to the common segmentation learned through a flat semantic taxonomy, we demonstrate that a hierarchical semantic taxonomy also in turn improves image segmentation.

2 Related Work

Hierarchical classification problem presents a unique challenge: the image remains the same, but the output changes in the semantic (text) space ("bird" - "Green Hermit"). Due to this formulation, prior research has primarily focused on embedding data into the **semantic (text) space**. For example, approaches include using additional loss functions (Bertinetto et al., 2020; Zeng et al., 2022) or representing entire taxonomies as flat, lengthy text inputs, as in BIOCLIP (Stevens et al., 2024). *In contrast*, our work takes a novel approach by addressing the problem from the perspective of the **visual space**. Specifically, we explore how hierarchical classification can be connected to visual grounding, examining how images can be analyzed at varying levels of detail—either more fine-grained or more holistic. **This visual-grounding perspective is unique and has not been explored in previous works.**

Prior hierarchical classification works can be categorized into three approaches: 1) flat classification (bottom-up) approach, 2) local classifier (top-down) approach, and 3) global classifier (multi-granularity) approach (Silla & Freitas, 2011).

1) The flat classification approach focuses on predicting fine-grained classes (e.g., leaf nodes) by leveraging taxonomy (Deng et al., 2014; Bertinetto et al., 2020; Karthik et al., 2021; Zhang et al., 2022; Zeng et al., 2022; Garg et al., 2022; Stevens et al., 2024). This bottom-up approach infers coarse classes from fine-grained predictions. While effective on clear and detailed images, it faces challenges in real-world scenarios where fine-grained predictions are challenging (e.g., birds at high altitudes). To address this, we propose a model that predicts across the entire taxonomy, which we believe provides greater robustness in practical applications.

2) The local classifier (top-down) approach leverages local information, such as higher-level class predictions, to make predictions at the next level. This design allows predictions at arbitrary nodes by stopping the inference process when a certain decision threshold is met, leading to more reliable predictions at higher levels (Deng et al., 2010; Wu et al., 2020; Brust & Denzler, 2019). As a result, these methods emphasize metrics such as the correctness-specificity trade-off (Valmadre, 2022). However, a disadvantage of this top-down approach is the propagation of errors from higher-level predictions to lower levels.

3) The global classifier (multi-granularity) approach aims to predict the entire taxonomy *at once*, unlike prior approaches. Most popular and effective methods uses a shared backbone with separate branches for each level (Zhu & Bain, 2017; Wehrmann et al., 2018; Chang et al., 2021; Liu et al., 2022; Chen et al., 2022; Jiang et al., 2024; Zhang et al., 2024). A critical challenge in this approach is maintaining *consistency* with the taxonomy in the predicted labels. To address this, Wang et al. (2023a) proposed a consistency-aware method by adjusting prediction scores through coarse-to-fine deduction and fine-to-coarse induction. However, we observed that using separate branches can lead to inconsistency, as each branch processes the image independently. To address this, we propose a model based on consistent visual grounding. To the best of our knowledge, no prior work has utilized visual segments to resolve inconsistency in hierarchical classification.

Unsupevised/Weakly-supervised Semantic Segmentation aims to group pixels without pixel-level annotations or using only class labels (Hwang et al., 2019; Ouali et al., 2020; Ke et al., 2022; 2024). These works employ hierarchical grouping to achieve meaningful segmentation *without* pixel-level labels. Here, "hierarchical" refers to part-to-whole visual grouping, where smaller units (e.g., a person's face or arm) are grouped into larger regions (e.g., the whole body). Based on our intuition that fine-grained classifiers need more detailed information, while coarse classifiers focus on broader groupings, our approach leverages these varying types of visual grouping. To implement this, we adopt the recently proposed CAST (Ke et al., 2024), whose graph pooling naturally supports consistent visual grouping. Notably, our work introduces the novel insight that part-to-whole spatial granularity can align with taxonomy hierarchies (e.g., finer segments for fine-grained labels, coarser segments for coarse labels), a connection not previously explored.

More detailed related work, including **Hierarchical Semantic Segmentation** can be found in Appendix B.

3 Consistent Hierarchical Classification

Our goal is to enhance the consistency of hierarchical recognition, thereby concurrently improving the accuracy of the model. To this end, we design a progressive learning scheme for hierarchical recognition, where the learning of each level contributes to the learning of the next level, instead of training separate models focusing on each individual level. Specifically, we address two types of inconsistency in hierarchical recognition. One is *visual inconsistency*, where classifiers at different levels attend to different regions (Figure 2). To address this, we propose H-CAST in Section 3.1. The other is *semantic inconsistency*, where predictions at different levels are not aligned within the taxonomy (e.g., "Plant"-"Hummingbird"). For this, we propose a new Tree-path KL Divergence loss that encodes parent-child relations to handle semantic inconsistency in Section 3.2. Figure 3 provides an overview of our method.

3.1 H-CAST for Visual Consistency

The areas of focus within the image need to differ when conducting classification at the fine-grained level compared to the coarse level. When distinguishing between similar-looking species (e.g., "Green Hermit" vs. "Ruby-throated Hummingbird"), the fine-grained recognition requires attention to *fine details* like the bird's beak and wings; meanwhile, at the coarse level (e.g., "bird" vs. "plant"), the attention shifts to *larger parts* such as the overall body of the bird. However, this shift in focus towards larger objects does not imply a sudden disregard for the previously focused details and a search for new larger objects. Rather, a natural approach involves combining detailed features such as the bird's beak, belly, and wings for accurate bird recognition. Therefore, we argue that **the hierarchical model should be grounded in consistent visual cues**. From this insight, we design a model where the details learned at the fine level (e.g., bird's beak and wings) are transferred to the coarse level as broader parts (e.g., bird's body) through consistent feature grouping.

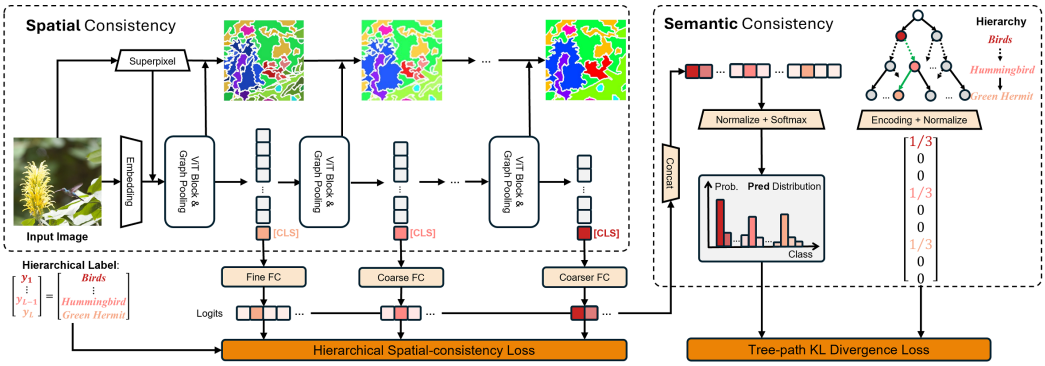


Figure 3: **Our method consists of two parts: Visual Consistency and Semantic Consistency module.** In the **Visual Consistency** module, the parsed images using superpixels are grouped based on related parts as they transition from fine to coarse levels. This guides that each hierarchical classifier focuses on the same corresponding regions while capturing different details of granularity. In the **Semantic Consistency** module, we incorporate hierarchical relationships between labels. This approach allows us to achieve consistent learning across the entire hierarchy. By promoting consistency, our method encourages classifiers at different levels to enhance overall performance, rather than conflicting with each other.

For internally consistent feature grouping, we build upon recent work CAST (Ke et al., 2024). CAST develops a hierarchical segmentation from fine to coarse, an internal part of the recognition process. However, their segmentation is driven by a flat recognition objective at the very end of visual parsing. We extend it by imposing fine-to-coarse semantic classification losses at different stages of segmentations throughout the visual parsing process. Our design reflects the intuition that finer segments can be helpful in capturing fine-grained details (e.g., beaks and wings) required for fine-grained recognition, whereas coarser segments can be effective in representing broader features (e.g., the body of a bird) needed for coarse-grained recognition. We have a single hierarchical recognition grounded on internally consistent segmentations, each driven by a classification objective at a certain granuality. We refer to our method as *Hierarchical-CAST* (H-CAST).

Consider a hierarchical recognition task where x denotes an image associated with hierarchical labels $y_1, \dots y_L$, encompassing a total of L levels in the hierarchy. Level L is the finest level (*i.e.,* leaf node), and Level 1 is the coarsest level (*i.e.,* root node). Then, given an image x, the hierarchical image recognition task is to predict labels at all levels across the hierarchy.

Let Z_l and S_l denote the feature and segments at l-th hierarchical level, respectively, Then, we obtain superpixels for image x by using the off-the-shelf algorithm SEEDS (Van den Bergh et al., 2012) to divide the image into regions with similar colors and local connectivity. These superpixels serve as input for the Vision Transformer (ViT) instead of fixed-size patches and simultaneously become the finest (initial) segments, S_{L+1}. l-th feature tokens Z_l is the concatenation of class tokens (Z_l^{class}) and segment tokens (Z_l^{seg}). Then, Graph pooling (Ke et al., 2024) aggregates segments with high feature similarity, allowing feature Z_l to progressively learn a more global visual context as it transitions from Z_L to Z_1.

For hierarchical recognition, we add a classification head (f_l) consisting of a single linear layer at each level. Then, we define the Hierarchical Visual-consistency loss as the sum of L cross-entropy losses (L_{CE}), denoted as

$$L_{HV} = \sum_{l=1}^{L} L_{CE}(f_l(Z_l^{class}), y_l). \tag{1}$$

Our approach differs from CAST in that while CAST uses the class token as the final objective, we design our model to incorporate hierarchical supervision during the training process. This ensures that labels from different levels progressively contribute to each other. In the Experimental section 4.5, we will demonstrate the effectiveness of our design, compared to alternative designs, including hierarchy supervision in the coarse-to-fine direction.

3.2 Tree-path KL Divergence loss for Semantic Consistency

To improve semantic consistency, we propose a new loss function called Tree-path KL Divergence loss, which directly incorporates hierarchical relationships between labels. Our idea is to encode the entire hierarchical structure so that a model can learn the hierarchy by outputting the tree of hierarchy. To this end, we first concatenate labels from all levels to create a distribution, as $Y = \frac{1}{L}[1_{y_L}; \ldots; 1_{y_1}]$, where 1_{y_l} represents the one-hot encoding for level l. Next, we concatenate the outputs of each classification head and then apply the log softmax function (LogSoftmax). We use Kullback–Leibler divergence loss (KL) to align this output with the ground truth distribution Y. Then, TK loss is calculated as follows.

$$L_{TK} = KL(\text{LogSoftmax}([f_L(Z_L^{class}); \ldots; f_1(Z_1^{class})]), Y). \tag{2}$$

This loss penalizes predictions that do not align with the taxonomy by simultaneously training on multiple labels within the hierarchy. Therefore, despite the simplicity, TK loss enables the model to enhance semantic consistency through this vertical encoding from the root (parent) node of the hierarchy level to the leaf (children) node. Our final loss becomes as follows, where α is a hyperparameter to control the weight of L_{TK},

$$L = L_{HV} + \alpha L_{TK}. \tag{3}$$

4 Experiments

We first demonstrate that hierarchical classification is a challenging problem that cannot be easily solved by vision foundation models, which also experience inconsistent predictions. Next, we compare our method against existing approaches and flat-level baselines on hierarchical classification benchmark datasets, showing that our approach significantly outperforms them. In addition, we justify the design of our method through ablation studies. Finally, we show that hierarchical supervision can surprisingly improve semantic segmentation as well.

4.1 Experimental Settings

Datasets. We use three widely used benchmarks in hierarchical recognition: BREEDS (Santurkar et al., 2021), CUB-200-2011 (Welinder et al., 2010), and FGVC-Aircraft (Maji et al., 2013). BREEDS, a subset of ImageNet (Russakovsky et al., 2015), includes four 2-level hierarchy datasets with different depths/parts based on the WordNet (Miller, 1995) hierarchy: Living-17, Non-Living-26, Entity-13, Entity-30. For BREEDS, we conduct training and validation using their source splits. BREEDS provide a wider class variety and larger sample size than CUB-200-2011 and FGVC-Aircraft, making it better suited for evaluating generalization performance. CUB-200-2011 comprises a 3-level hierarchy with order, family, and species; FGVC-Aircraft consists of a 3-level hierarchy including maker, family, and model (*e.g.*, Boeing - Boeing 707 - 707-320); Table 1 provides a description of the datasets.

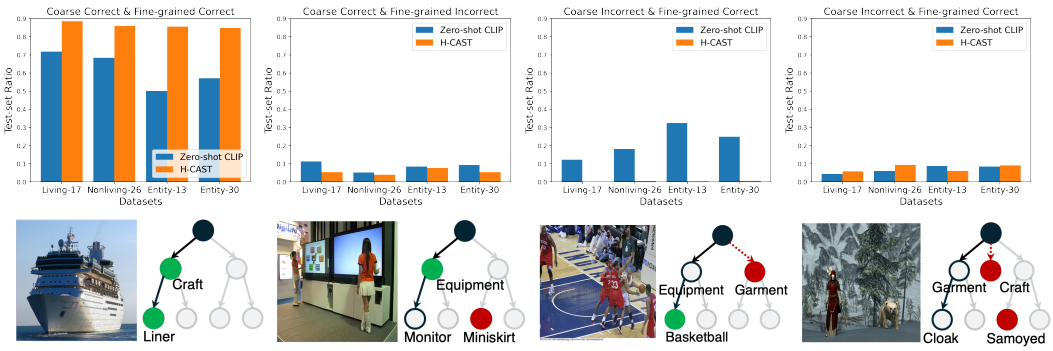
Table 1: **Benchmark Datasets.**

Datasets	L-17	NL-26	E-13	E-30	CUB	Aircraft
# Levels	2	2	2	2	3	3
# of classes	17-34	26-52	13-130	30-120	13-38-200	30-70-100
# Train images	44.2K	65.7K	167K	154K	5,944	6,667
# Test images	1.7K	2.6K	6.5K	6K	5,794	3,333

Evaluation Metrics. We evaluate our models using metrics for both accuracy and consistency.

- ***level-accuracy***: the proportion of correctly classified instances at each level (Chang et al., 2021).
- ***weighted average precision (wAP)*** (Liu et al., 2022): $\text{wAP} = \sum_{l=1}^{L} \frac{N_l}{\sum_{k=1}^{L} N_k} P_l$, where N_l and P_l denote the number of classes and Top-1 classification accuracy at level l, respectively. This metric considers the classification difficulty across different hierarchies.
- ***Tree-based InConsistency Error rate (TICE)*** (Wang et al., 2023a): $\text{TICE} = n_{\text{ic}}/N$, where n_{ic} denotes the number of samples with inconsistent prediction paths, and N refers to the number of all test samples. This tests whether the prediction path exists in the tree (consistency).
- ***Full-Path Accuracy (FPA)***: $\text{FPA} = n_{\text{ac}}/N$, where n_{ac} refers to the number of samples with all level of labels correctly predicted. This metric evaluates **both accuracy and consistency**, **ultimately representing our primary metric of interest**.

The difference between FPA and TICE is illustrated in Table 8.



(a) Coarse O & Fine O (b) Coarse O & Fine X (c) Coarse X & Fine O (d) Coarse X & Fine X

Figure 4: **Consistent predictions for hierarchical classification cannot be easily solved using a vision foundation model.** We evaluate the performance of CLIP (Radford et al., 2021) on the 2-level BREEDS dataset (top) and provide misclassification examples from Entity-13 for each case (bottom). (a) CLIP struggles to make consistent and correct predictions, achieving about 50% accuracy on the Entity-13 dataset. (c) Notably, CLIP has difficulty with coarse predictions with broader concepts. In contrast, our H-CAST accurately predicts in all cases.

Comparison methods. First, we evaluate our H-CAST with representative models in hierarchical classification, **FGN** (Chang et al., 2021) and **HRN** (Chen et al., 2022). **FGN** uses level-specific heads to avoid negative transfer across granularity levels, while **HRN** employs residual connections to capture label relationships and a hierarchy-based probabilistic loss. As no existing multi-granularity methods use a vision transformer (ViT) backbone, we include **Hier-ViT**, a ViT-based model without consistent visual grounding. Similar to our approach, **Hier-ViT** trains each hierarchy level using the class token from the last l blocks. To establish a ceiling baseline, we compare with flat models trained at a single hierarchy level. **Flat-ViT** classifies one level using the ViT class token, while **Flat-CAST** trains independent models for each level using the CAST architecture (Ke et al., 2024). We also compare with **HiE** (Jain et al., 2023), which improves fine-grained predictions via post-hoc correction using a coarse model. Note that flat models require separate models for each hierarchy level, leading to increased storage and training costs.

Architecture and Training. For baseline methods, we use their codebase for optimal hyperparameters. For a fair comparison, we use ViT-Small and CAST-Small models of corresponding sizes. As in CAST, segmentation granularity is set to 64, 32, 16, 8 after 3, 3, 3, 2 encoder blocks, respectively. Our training progresses from fine to coarse levels, with each segment corresponding accordingly. The initial number of superpixels is set to 196, and all data is trained with a batch size of 256 for 100 epochs. Following the literature, we use ImageNet pre-trained models for the Aircraft and CUB datasets. For the ImageNet subset BREEDS dataset, we train the models from scratch. Detailed hyperparameter settings can be found in the Appendix C.

4.2 Hierarchical Classification with Vision Foundation Models

First, to demonstrate that longstanding hierarchical classification is not easily solved by today's vision foundation models, we evaluate CLIP (Radford et al., 2021)'s performance on the 2-level BREEDS dataset. The top row of Figure 4 shows prediction rates on the test set, while the bottom row presents examples from the Entity-13 dataset. Even considering the zero-shot prediction, Figure 4 (a) shows that the overall ratio of correct predictions for both coarse and fine-grained classifications is low, with only around 50% accuracy on the Entity-13 dataset. Figure 4 (c) further highlights significant errors in coarse predictions when addressing broader concepts. This indicates that hierarchical classification remains a challenging problem, even in the era of vision foundation models. Furthermore, when we examine the misclassification examples in bottom row of Figure 4, we can see that CLIP focuses on different object areas for coarse and fine-grained predictions. For example, in (b), CLIP predicts "Equipment" in the coarse prediction but predicts "Miniskirt" in the fine-grained prediction instead of "Monitor" (a child of Equipment). However, our model, based on consistent visual segments, can make correct predictions in all cases.

Table 2: **Our H-CAST outperforms both hierarchical baselines (FGN (Chang et al., 2021), HRN (Chen et al., 2022), Hier-ViT) and flat baselines (ViT (Dosovitskiy et al., 2020), CAST (Ke et al., 2024), HiE (Jain et al., 2023)) on BREEDS.** It achieves a 4.3-6.4 percentage point gain in FPA metric over HRN with significantly fewer parameters. Additionally, H-CAST surpasses Hier-ViT by over 11 percentage points, demonstrating that its success is due to our consistent visual grounding and Tree-path Loss, rather than merely adding hierarchy supervision to a ViT backbone. (Higher the metric is the best, except TICE.) 'ViT-S' refers to ViT-Small, while 'RN-50' denotes ResNet-50.

					Living-17 (17-34)					Non-Living-26 (26-52)				
		Backbone	# Params	Batch size	FPA	Coarse	Fine	wAP	TICE	FPA	Coarse	Fine	wAP	TICE
Flat	Flat-ViT	ViT-S	65.0M	256	66.24	75.71	72.06	73.28	17.11	57.46	67.50	65.73	57.46	23.27
	Flat-ViT + HiE	ViT-S	65.0M	256	67.59	75.71	71.35	72.81	9.88	59.73	67.50	65.31	66.04	13.69
	Flat-CAST	ViT-S	78.5M	256	78.82	88.06	82.88	84.61	8.82	76.17	84.77	81.08	82.31	11.77
	Flat-CAST + HiE	ViT-S	78.5M	256	81.59	88.06	83.24	84.85	5.18	79.23	84.77	81.39	82.51	6.19
Hierarchy	FGN	RN-50	24.8M	128	63.82	72.59	68.00	69.53	12.12	60.81	69.46	65.77	67.00	16.46
	HRN	RN-50	70.8M	128	50.59	62.53	56.76	58.69	19.94	45.42	54.81	52.58	53.32	25.73
	HRN	RN-50	70.8M	8 [1]	79.18	87.53	81.47	83.49	6.29	76.31	82.38	80.15	80.90	9.54
	Hier-ViT	ViT-S	21.7M	256	74.06	80.94	74.88	76.90	10.50	72.04	73.31	58.39	70.03	12.45
	Ours (H-CAST)	ViT-S	26.2M	256	**85.12**	**90.82**	**85.24**	**87.10**	**3.19**	**82.67**	**87.89**	**83.15**	**84.73**	**5.26**
	Our Gains over SOTA				+5.94	+3.29	+3.77	+3.61	+3.10	+6.36	+5.51	+3.00	+3.83	+4.28

					Entity-13 (13-130)					Entity-30 (30-120)				
		Backbone	# Params	Batch size	FPA	Coarse	Fine	wAP	TICE	FPA	Coarse	Fine	wAP	TICE
Flat	Flat-ViT	ViT-S	65.0M	256	64.22	76.28	76.06	76.08	21.33	66.93	76.28	74.35	74.77	18.75
	Flat-ViT + HiE	ViT-S	65.0M	256	65.20	76.47	74.91	75.05	15.68	68.77	76.47	73.92	74.43	11.08
	Flat-CAST	ViT-S	78.5M	256	78.63	87.80	83.72	84.09	10.65	82.67	87.89	83.15	84.73	5.26
	Flat-CAST + HiE	ViT-S	78.5M	256	79.52	87.80	83.40	83.80	6.83	83.70	87.89	84.30	85.02	4.20
Hierarchy	FGN	RN-50	24.8M	128	74.23	85.35	78.00	78.67	9.43	68.52	77.47	73.18	74.04	13.62
	HRN	RN-50	70.8M	128	52.62	71.43	59.49	60.58	20.05	46.17	58.35	55.07	55.72	28.20
	HRN	RN-50	70.8M	8	81.43	90.00	84.48	84.98	6.34	79.85	86.57	83.35	83.99	8.38
	Hier-ViT	ViT-S	21.7M	256	74.63	86.95	75.39	77.70	5.19	73.01	81.38	74.10	74.76	11.61
	Ours (H-CAST)	ViT-S	26.2M	256	**85.68**	**93.42**	**86.15**	**87.60**	**1.69**	**84.83**	**90.23**	**85.45**	**85.88**	**2.57**
	Our Gains over SOTA				+4.25	+3.42	+1.67	+2.62	+4.65	+4.98	+3.66	+2.10	+1.89	+5.81

4.3 Consistent Hierarchical Classification on Benchmarks

Table 2 shows the comparison with baselines on BREEDS dataset. Our H-CAST outperforms not only hierarchical classification baselines like FGN, HRN, and Hier-ViT, but also flat baselines such as ViT, CAST, and HiE. Notably, it achieves a 4.3-6.4 percentage point gain on the FPA metric, compared to the HRN, despite using significantly fewer parameters. Notably, H-CAST surpasses Hier-ViT by over 11 percentage points, demonstrating that its success can be attributed to our consistent visual grounding and Tree-path Loss, rather than simplly applying hierarchy supervision to a ViT backbone.

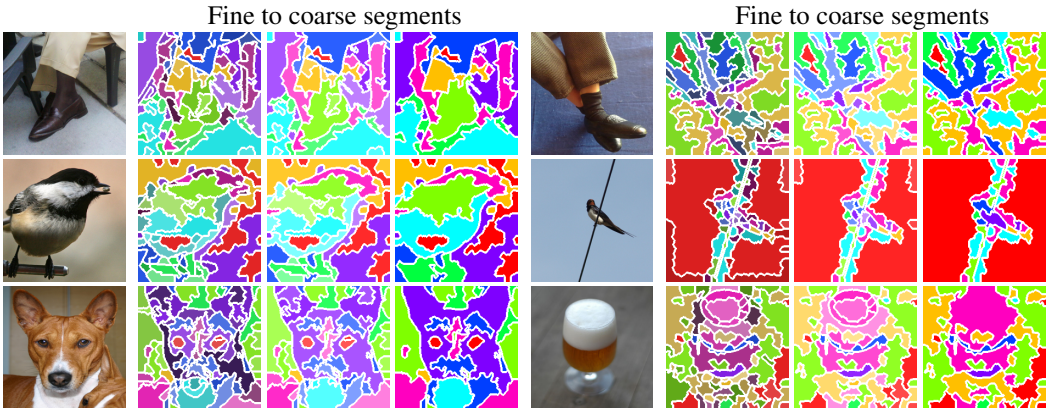
In addition, the visualization of attention maps shows that H-CAST's lower blocks focus on detailed and localized regions. In contrast, the upper blocks attend to broader areas, supporting consistent visual grounding for hierarchical classification. We show the results in Appendix D.2.

It should be noted that flat models train a separate model for each hierarchy level, leading to substantially greater memory and training time requirements. The superior performance of our model compared to these flat models demonstrates its effectiveness in hierarchical classification. Similar results are also observed in the Aircraft and CUB datasets, which are included in Appendix D.4.

4.4 Enhancing Spatial Focus and Interpretability with H-CAST

H-CAST can guide consistent spatial focus and improve the interpretability of the model's predictions. To explore how our model learns segments with varying granularity from fine to coarse levels, we visualize segments based on hierarchy levels on BREEDS Entity-30 dataset in Figure 5. In full-path correct prediction cases, where predictions at all levels are correct (Figure 5, Left), visual details are effectively captured at the fine level and consistently grouped to identify larger objects at the coarse level. However, in full-path incorrect prediction cases, where predictions at all levels are incorrect (Figure 5, Right), the model's segments fail to recognize the object accurately. This underscores an added benefit of incorporating segments in hierarchical recognition models, as it not only contributes to consistent predictions but also enhances the *interpretability* of the model's predictions.

[1] HRN's performance is highly sensitive to input and batch size, especially requiring smaller batches, which doubled the training time compared to ours.



a) Full-Path Correct Predictions b) Full-Path Incorrect Predictions

Figure 5: **H-CAST enhances the interpretability of the model's predictions.** We visualize feature grouping from fine to coarse for full-path correct and incorrect predictions on the Entity-30 dataset. For full-path correct predictions (all levels correct), visual details are effectively grouped to identify larger objects at coarser levels. In contrast, for full-path incorrect predictions (all levels incorrect), segments fail to recognize the object. For example, in the first row, for the correctly predicted shoe image, the shoe and ankle parts are grouped together in green, showing coherent segmentation. In contrast, the incorrect prediction case shows highly fractured segments, failing to capture the grouping of the shoe. This demonstrates the added benefit of using segments in hierarchical recognition, providing both guidance on focus and improved interpretability.

4.5 Ablation Analysis of Architecture Design and Loss Function in H-CAST

Fine-to-Coarse vs. Coarse-to-Fine learning. Our model adopts a Fine-to-Coarse learning strategy, where it first learns fine labels in the lower block and progressively integrates coarser labels. This contrasts with conventional ResNet-based models, which typically learn coarse features first (Zhu & Bain, 2017; Yan et al., 2015; Zeiler & Fergus, 2014). To evaluate the effectiveness of our Fine-to-Coarse architecture, we compare it with two baselines.

The first baseline, Coarse-to-Fine, learns coarse labels in the lower block and fine labels in the upper block, following a conventional hierarchy. The second, Fine-Coarse Merging, combines the class token from the lower block with coarser segments from the upper block for coarse labels and vice versa for fine labels. This approach intuitively leverages features of varying granularities. For fairness, we do not use Tree-path KL Divergence loss in these comparisons. Table 3 summarizes the results on the FGVC-Aircraft dataset. The results show that Coarse-to-Fine learning shows the lowest fine-grained accuracy. Fine-Coarse Merging achieves slightly higher fine-grained accuracy but with a significant increase in parameters due to the use of segment features in the classification head. Our Fine-to-Coarse strategy balances simplicity and strong performance, making it an effective choice for hierarchical classification.

Table 3: **Coarse-to-Fine learning scheme achieves best overall performance.** We report the impact of hierarchical learning direction on FGVC-Aircraft.

Learning Direction	FPA	maker	family	model	wAP
Coarse-to-Fine	82.01	93.16	89.92	84.10	87.50
Fine-Coarse merging	81.76	93.52	**90.31**	**84.58**	**87.93**
Fine-to-Coarse	**82.66**	**94.27**	90.19	84.40	87.91

Ablation Studies on the Proposed Losses. To evaluate the effectiveness of the proposed loss, we conduct two ablation studies. First, we assess the individual contributions of Hierarchical Spatial-consistency loss L_{HS} and Tree-path KL Divergence loss L_{TK} on Aircraft. Table 4 shows that both losses significantly enhance performance, with their combined use achieving the best accuracy and consistency.

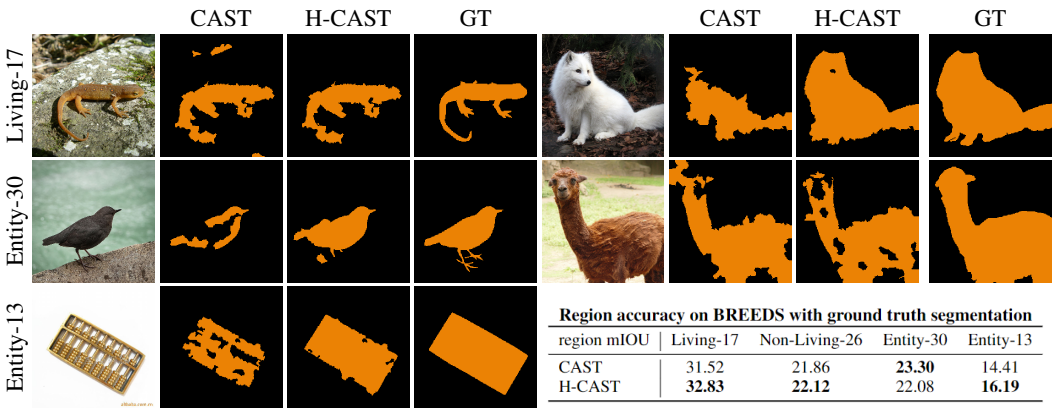
Next, to examine how the choice of loss function affects performance, we replace our proposed KL Divergence loss with two alternative losses: Binary Cross Entropy (BCE) loss and Flat Consistency loss. BCE loss directly substitutes the KL divergence component in our setup. Flat Consistency loss, inspired by a bottom-up approach, infers coarse predictions from fine-grained predictions, using BCE

Table 4: **Utilizing both losses yields best performance on Aircraft.**

L_{HS}	L_{TK}	FPA	maker	family	model	wAP
✗	✓	82.48	94.30	90.37	84.04	87.80
✓	✗	82.66	94.27	90.19	84.40	87.91
✓	✓	**83.72**	**94.96**	**91.39**	**85.33**	**88.90**

Table 5: **KL Div. loss shows best performance on Aircraft.**

Sem. Consis.	FPA	maker	family	model	wAP
Flat Consis.	82.87	94.63	90.94	84.97	88.51
BCE	82.18	94.21	90.13	84.88	88.11
KL Div. Loss	**83.72**	**94.96**	**91.39**	**85.33**	**88.90**



Region accuracy on BREEDS with ground truth segmentation

region mIOU	Living-17	Non-Living-26	Entity-30	Entity-13
CAST	31.52	21.86	**23.30**	14.41
H-CAST	**32.83**	**22.12**	22.08	**16.19**

Figure 6: **H-CAST improves segmentation by leveraging hierarchical taxonomy.** We visualize segmentation results on BREEDS dataset and measure the region mIOU of fine-level objects for samples with segmentation ground truth (GT) from ImageNet-S (Gao et al., 2022). In the visualized images, we can observe that H-CAST better captures the overall shape in a more coherent manner compared to CAST. In quantitative evaluation, H-CAST outperforms CAST in most datasets despite using coarse-level supervision for the last-level segments whereas CAST employs fine-level supervision. It is surprising to find that the taxonomy hierarchy can help part-to-whole segmentation.

to match them with the ground truth. As shown in Table 5 on Aircraft dataset, KL Divergence loss achieves the best FPA, demonstrating superior accuracy and consistency. Additional ablation results on Living-17 are provided in the Appendix D.3.

4.6 Additional Benefits of Hierarchical Classification for Segmentation

Hierarchical semantic recognition enhances segmentation. Although our primary focus is hierarchical recognition, we investigate whether incorporating hierarchical label information can also improve segmentation. Figure and Table 6 provide a qualitative and quantitative comparison between H-CAST and CAST. H-CAST, which uses varying granularity supervision for segments, outperforms CAST, which employs fine-grained level supervision, on most datasets such as Living-17, Non-Living-26, and Entity-13. The visualized results show that H-CAST better captures the overall shape in a more coherent manner compared to CAST. These findings demonstrate that utilizing hierarchical taxonomy benefits not only recognition but also segmentation. Details of the evaluation method and more visualization comparison with CAST are included in the Appendix D.6 and Figure 9.

5 Conclusion and Discussions

In this work, we tackle the challenge of inconsistent predictions in hierarchical classification by introducing consistent visual grounding. By leveraging varying granularity segments, our approach guides hierarchical classifiers to focus on coherent and relevant regions across levels, guiding alignment between coarse and fine-grained predictions. We demonstrates its effectiveness across benchmark datasets. While our method shows strong performance, it currently faces limitations in scaling to deep hierarchies with 10–20 levels. Extending the approach to handle such hierarchies and exploring its scalability in highly imbalanced taxonomies are promising directions for future research. Additionally, the computational overhead from superpixel generation and graph pooling increases processing time, and optimizing these components will be an important step toward improving efficiency.

REFERENCES

Luca Bertinetto, Romain Mueller, Konstantinos Tertikas, Sina Samangooei, and Nicholas A Lord. Making better mistakes: Leveraging class hierarchies with deep networks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2020.

Clemens-Alexander Brust and Joachim Denzler. Integrating domain knowledge: using hierarchies to improve deep classifiers. In *Asian conference on pattern recognition*. Springer, 2019.

Dongliang Chang, Kaiyue Pang, Yixiao Zheng, Zhanyu Ma, Yi-Zhe Song, and Jun Guo. Your" flamingo" is my" bird": fine-grained, or not. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021.

Jingzhou Chen, Peng Wang, Jian Liu, and Yuntao Qian. Label relation graphs enhanced hierarchical residual network for hierarchical multi-granularity classification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022.

Ekin D Cubuk, Barret Zoph, Jonathon Shlens, and Quoc V Le. Randaugment: Practical automated data augmentation with a reduced search space. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition workshops*, 2020.

Jia Deng, Alexander C Berg, Kai Li, and Li Fei-Fei. What does classifying more than 10,000 image categories tell us? In *Computer Vision–ECCV 2010: 11th European Conference on Computer Vision, Heraklion, Crete, Greece, September 5-11, 2010, Proceedings, Part V 11*. Springer, 2010.

Jia Deng, Nan Ding, Yangqing Jia, Andrea Frome, Kevin Murphy, Samy Bengio, Yuan Li, Hartmut Neven, and Hartwig Adam. Large-scale object classification using label relation graphs. In *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part I 13*, 2014.

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.

Shanghua Gao, Zhong-Yu Li, Ming-Hsuan Yang, Ming-Ming Cheng, Junwei Han, and Philip Torr. Large-scale unsupervised semantic segmentation. *TPAMI*, 2022.

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

Ju He, Jieneng Chen, Ming-Xian Lin, Qihang Yu, and Alan L Yuille. Compositor: Bottom-up clustering and compositing for robust part and object segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023.

Jyh-Jing Hwang, Stella X Yu, Jianbo Shi, Maxwell D Collins, Tien-Ju Yang, Xiao Zhang, and Liang-Chieh Chen. Segsort: Segmentation by discriminative sorting of segments. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2019.

Kanishk Jain, Shyamgopal Karthik, and Vineet Gandhi. Test-time amendment with a coarse classifier for fine-grained classification. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023.

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

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.

Tsung-Wei Ke, Jyh-Jing Hwang, Yunhui Guo, Xudong Wang, and Stella X Yu. Unsupervised hierarchical semantic segmentation with multiview cosegmentation and clustering transformers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022.

Tsung-Wei Ke, Sangwoo Mo, and Stella X. Yu. Learning hierarchical image segmentation for recognition and by recognition. In *The Twelfth International Conference on Learning Representations*, 2024.

Liulei Li, Tianfei Zhou, Wenguan Wang, Jianwu Li, and Yi Yang. Deep hierarchical semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022.

Liulei Li, Wenguan Wang, and Yi Yang. Logicseg: Parsing visual semantics with neural logic learning and reasoning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023.

Feng Liang, Bichen Wu, Xiaoliang Dai, Kunpeng Li, Yinan Zhao, Hang Zhang, Peizhao Zhang, Peter Vajda, and Diana Marculescu. Open-vocabulary semantic segmentation with mask-adapted clip. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023.

Yang Liu, Lei Zhou, Pengcheng Zhang, Xiao Bai, Lin Gu, Xiaohan Yu, Jun Zhou, and Edwin R Hancock. Where to focus: Investigating hierarchical attention relationship for fine-grained visual classification. In *European Conference on Computer Vision*. Springer, 2022.

S. Maji, J. Kannala, E. Rahtu, M. Blaschko, and A. Vedaldi. Fine-grained visual classification of aircraft. Technical report, 2013.

George A Miller. Wordnet: a lexical database for english. *Communications of the ACM*, 1995.

Yassine Ouali, Céline Hudelot, and Myriam Tami. Autoregressive unsupervised image segmentation. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part VII 16*. Springer, 2020.

Lu Qi, Jason Kuen, Weidong Guo, Jiuxiang Gu, Zhe Lin, Bo Du, Yu Xu, and Ming-Hsuan Yang. Aims: All-inclusive multi-level segmentation for anything. *Advances in Neural Information Processing Systems*, 2024.

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.

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. *International Journal of Computer Vision (IJCV)*, 2015.

Shibani Santurkar, Dimitris Tsipras, and Aleksander Madry. {BREEDS}: Benchmarks for subpopulation shift. In *International Conference on Learning Representations*, 2021.

Ramprasaath R Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *Proceedings of the IEEE international conference on computer vision*, 2017.

Carlos N Silla and Alex A Freitas. A survey of hierarchical classification across different application domains. *Data mining and knowledge discovery*, 2011.

Rishubh Singh, Pranav Gupta, Pradeep Shenoy, and Ravikiran Sarvadevabhatla. Float: Factorized learning of object attributes for improved multi-object multi-part scene parsing. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 1445–1455, 2022.

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 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024.

Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jon Shlens, and Zbigniew Wojna. Rethinking the inception architecture for computer vision. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016.

Jack Valmadre. Hierarchical classification at multiple operating points. *Advances in Neural Information Processing Systems*, 2022.

Michael Van den Bergh, Xavier Boix, Gemma Roig, Benjamin De Capitani, and Luc Van Gool. Seeds: Superpixels extracted via energy-driven sampling. In *Computer Vision–ECCV 2012: 12th European Conference on Computer Vision, Florence, Italy, October 7-13, 2012, Proceedings, Part VII 12*, 2012.

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*, 2023a.

Wenhao Wang, Yifan Sun, Wei Li, and Yi Yang. TransHP: Image classification with hierarchical prompting. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023b.

Xudong Wang, Shufan Li, Konstantinos Kallidromitis, Yusuke Kato, Kazuki Kozuka, and Trevor Darrell. Hierarchical open-vocabulary universal image segmentation. *Advances in Neural Information Processing Systems*, 2024.

Jonatas Wehrmann, Ricardo Cerri, and Rodrigo Barros. Hierarchical multi-label classification networks. In *International conference on machine learning*. PMLR, 2018.

P. Welinder, S. Branson, T. Mita, C. Wah, F. Schroff, S. Belongie, and P. Perona. Caltech-UCSD Birds 200. Technical Report CNS-TR-2010-001, California Institute of Technology, 2010.

Tz-Ying Wu, Pedro Morgado, Pei Wang, Chih-Hui Ho, and Nuno Vasconcelos. Solving long-tailed recognition with deep realistic taxonomic classifier. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part VIII 16*, 2020.

Zhicheng Yan, Hao Zhang, Robinson Piramuthu, Vignesh Jagadeesh, Dennis DeCoste, Wei Di, and Yizhou Yu. Hd-cnn: hierarchical deep convolutional neural networks for large scale visual recognition. In *Proceedings of the IEEE international conference on computer vision*, 2015.

Sangdoo Yun, Dongyoon Han, Seong Joon Oh, Sanghyuk Chun, Junsuk Choe, and Youngjoon Yoo. Cutmix: Regularization strategy to train strong classifiers with localizable features. In *Proceedings of the IEEE/CVF international conference on computer vision*, 2019.

Matthew D Zeiler and Rob Fergus. Visualizing and understanding convolutional networks. In *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part I 13*, 2014.

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.

Hongyi Zhang, Moustapha Cisse, Yann N Dauphin, and David Lopez-Paz. mixup: Beyond empirical risk minimization. *arXiv preprint arXiv:1710.09412*, 2017.

Shichuan Zhang, Sunyi Zheng, Zhongyi Shui, and Lin Yang. Hls-fgvc: Hierarchical label semantics enhanced fine-grained visual classification. In *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2024.

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.

Xinqi Zhu and Michael Bain. B-cnn: branch convolutional neural network for hierarchical classification. *arXiv preprint arXiv:1709.09890*, 2017.

Appendix

A Quantitative Evidence for Consistent Visual Grounding

To quantitatively validate our observation that inconsistent predictions often occur when coarse and fine-grained classifiers focus on different regions in Figure 2, we analyze the Grad-CAM (Selvaraju et al., 2017) heatmaps of these classifiers. Specifically, we compute two metrics: the overlap score and the correlation score.

The **overlap score** quantifies the degree to which the regions activated by the two classifiers coincide. For each heatmap, we define a significant region as the set of pixels where activation values exceed a threshold. Specifically, the overlap count (O) measures the number of overlapping pixels between heatmaps A and B, where both values exceed a threshold ($\tau = 0.001$). It is defined as:

$$O = \sum_{i,j} \left[M_A(i,j) \wedge M_B(i,j) \right], \tag{4}$$

where $M_A(i,j)$ and $M_B(i,j)$ are binary masks indicating significant regions in A and B, respectively. These masks are defined as:

$$M_A(i,j) = \begin{cases} 1 & \text{if } A(i,j) > \tau, \\ 0 & \text{otherwise,} \end{cases} \quad M_B(i,j) = \begin{cases} 1 & \text{if } B(i,j) > \tau, \\ 0 & \text{otherwise.} \end{cases} \tag{5}$$

The **correlation score** measures the linear relationship between the activation values of the overlapping regions in the two heatmaps. Let A_k and B_k the values in the overlapping regions, then correlation score is computed as:

$$R = \frac{\sum_{k=1}^{n}(A_k - \mu_A)(B_k - \mu_B)}{\sqrt{\sum_{k=1}^{n}(A_k - \mu_A)^2 \cdot \sum_{k=1}^{n}(B_k - \mu_B)^2}}, \tag{6}$$

where n is the number of overlapping pixels, μ_A is the mean of $\{A_k\}$, and μ_B is the mean of $\{B_k\}$.

Higher overlap and correlation scores indicate stronger agreement between the regions attended to by the two classifiers. Conversely, lower scores highlight a lack of alignment in their focus.

Interestingly, empirical results from the FGN model (Chang et al., 2021) on the Entity-30 dataset show that when both classifiers make correct predictions, the overlap and correlation scores are significantly higher. In contrast, incorrect predictions correspond to notably lower scores, as shown in Table 6. These findings support our hypothesis that aligning the focus of coarse- and fine-grained classifiers enhances both prediction accuracy and consistency.

Table 6: **Overlap and correlation scores between coarse and fine-grained Grad-CAM heatmaps.** This shows that correct predictions correspond to higher overlap and correlation between coarse and fine-grained classifiers, highlighting the importance of aligning classifier focus for accuracy and consistency.

Overlap		Fine-grained	
		True	False
Coarse	True	**0.51**± **0.20**	0.25 ± 0.13
	False	0.36 ± 0.18	0.37 ± 0.19

(a) Overlap Scores

Correlation		Fine-grained	
		True	False
Coarse	True	**0.70** ± **0.26**	-0.02 ± 0.40
	False	0.30 ± 0.42	0.35 ± 0.41

(b) Correlation Scores

B Additional Related Work

Hierarchical Classification can be categorized into three approaches: 1) flat classification (bottom-up) approach, 2) local classifier (top-down) approach, and 3) global classifier (multi-granularity) approach (Silla & Freitas, 2011).

1) The flat classification approach focuses on predicting fine-grained classes (e.g., leaf nodes) by leveraging taxonomy (Deng et al., 2014; Zhang et al., 2022; Zeng et al., 2022). It is often referred to as a bottom-up method because higher-level coarse classes can be inferred from the predicted fine-grained classes. Various methods have been proposed to effectively use hierarchical information. For example, hierarchical cross-entropy (HXE) loss (Bertinetto et al., 2020) reweights cross-entropy terms along the hierarchy tree based on class depth. Inspired by transformer prompting techniques, TransHP (Wang et al., 2023b) introduced coarse-class prompt tokens to improve fine-grained classification accuracy. Recently, BIOCLIP (Stevens et al., 2024), trained on large-scale Tree of Life data, achieved superior few-shot and zero-shot performance using a CLIP (Radford et al., 2021) contrastive objective on text combining fine-grained and higher-level classes. One of the actively studied topics is minimizing "mistake severity" (e.g., the tree distance between incorrect predictions and the ground truth) (Bertinetto et al., 2020; Karthik et al., 2021; Garg et al., 2022).

However, while effective on clear and detailed images, this approach struggles in real-world scenarios where fine-grained predictions are challenging (e.g., birds flying at high altitude), leading to incorrect predictions at higher levels. To address this, we propose a model that predicts across the entire taxonomy, which we believe provides greater robustness in practical applications.

2) The local classifier (top-down) approach leverages local information, such as higher-level class predictions, to make predictions at the next level. This design allows predictions at arbitrary nodes by stopping the inference process when a certain decision threshold is met, leading to more reliable predictions at higher levels (Deng et al., 2010; Wu et al., 2020; Brust & Denzler, 2019). As a result, these methods emphasize metrics such as the correctness-specificity trade-off (Valmadre, 2022). While a single model is commonly used, HiE (Jain et al., 2023) adjusts fine-level predictions post-hoc using coarse predictions from independently trained classifiers. However, a disadvantage of this top-down approach is the propagation of errors from higher-level predictions to lower levels.

3) The global classifier (multi-granularity) approach aims to predict the entire taxonomy *at once*, unlike prior approaches. Most popular and effective methods uses a shared backbone with separate branches for each level (Zhu & Bain, 2017; Wehrmann et al., 2018; Chang et al., 2021; Liu et al., 2022; Chen et al., 2022; Jiang et al., 2024; Zhang et al., 2024). The key difference lies in how the hierarchical relationships are modeled. For instance, in FGN (Chang et al., 2021), finer features are concatenated to predict coarse labels, whereas in HRN (Chen et al., 2022), coarse features are added to finer features through residual connections. A critical issue in this approach is maintaining *consistency* with the taxonomy in the predicted labels. To address this, Wang et al. (2023a) proposed a consistency-aware method by adjusting prediction scores through coarse-to-fine deduction and fine-to-coarse induction. However, we observed that using separate branches can lead to inconsistency, as each branch processes the image independently. To address this, we propose a model based on consistent visual grounding. To the best of our knowledge, no prior work has utilized visual segments to resolve inconsistency in hierarchical classification.

Hierarchical Semantic Segmentation aims to group and classify each pixel according to a class hierarchy (Li et al., 2022; Singh et al., 2022; Li et al., 2023; He et al., 2023; Wang et al., 2024; Qi et al., 2024), with pixel grouping varying based on the taxonomy used. However, these works require *pixel-level annotations*, which are not available in hierarchical classification. In addition, while these method focus on precise pixel-level grouping, our work leverage unsupervised segments of varying granularities within the image for hierarchical clasification.

Unsupevised/Weakly-supervised Semantic Segmentation aims to group pixels without pixel-level annotations or using only class labels (Hwang et al., 2019; Ouali et al., 2020; Ke et al., 2022; 2024). These works employ hierarchical grouping to achieve meaningful segmentation *without* pixel-level labels. Here, "hierarchical" refers to part-to-whole visual grouping, where smaller units (e.g., a person's face or arm) are grouped into larger regions (e.g., the whole body). Based on our intuition that fine-grained classifiers need more detailed information, while coarse classifiers focus on broader groupings, our approach leverages these varying types of visual grouping. To implement this, we

adopt the recently proposed CAST (Ke et al., 2024), whose graph pooling naturally supports consistent visual grouping. Notably, our work introduces the novel insight that part-to-whole segmentation can align with taxonomy hierarchies (e.g., finer segments for fine-grained labels, coarser segments for coarse labels), a connection not previously explored.

C HYPERPARAMETERS FOR TRAINING.

We show hyper-parameter settings in Table 7. Codes will be made publicly available.

Table 7: **Hyper-parameters for training H-CAST and ViT on FGVC-Aircraft, CUB-200-2011, and BREEDS.** We follow mostly the same set up as CAST (Ke et al., 2024).

Parameter	Aircraft	CUB	BREEDS
batch_size	256	256	256
crop_size	224	224	224
learning_rate	$1e^{-3}$	5^{e-4}	5^{e-4}
weight_decay	0.05	0.05	0.05
momentum	0.9	0.9	0.9
total_epochs	100	100	100
warmup_epochs	5	5	5
warmup_learning_rate	$1e^{-4}$	$1e^{-6}$	1^{e-6}
optimizer	Adam	Adam	Adam
learning_rate_policy	Cosine decay	Cosine decay	Cosine decay
augmentation (Cubuk et al., 2020)	RandAug(9, 0.5)	RandAug(9, 0.5)	RandAug(9, 0.5)
label_smoothing (Szegedy et al., 2016)	0.1	0.1	0.1
mixup (Zhang et al., 2017)	0.8	0.8	0.8
cutmix (Yun et al., 2019)	1.0	1.0	1.0
α (weight for TK loss)	0.5	0.5	0.5
ViT-S: # Tokens	$[196]_{\times 11}$		
CAST-S: # Tokens	$[196]_{\times 3}, [64]_{\times 3}, [32]_{\times 3}, [16]_{\times 2}$		

D Additional Experiments

D.1 Comparison between FPA and TICE.

FPA evaluates both accuracy and consistency, while TICE focuses solely on consistency. Achieving high FPA is the primary goal in hierarchical classification. The distinction between FPA and TICE is shown in Table 8.

Table 8: **FPA considers both correctness and consistency.** While TICE (Wang et al., 2023a) measures only consistency, FPA marks predictions as positive only when they are both correct and consistent.

								
TICE	✓	✓	✗	✗	✗	✗	✓	✓
FPA	✓	✗	✗	✗	✗	✗	✗	✗

D.2 Visualization of Attention Map

To validate our claim that the model guides classifiers toward consistent visual grounding, we visualize attention maps from H-CAST in Figure 7. The visualizations demonstrate that as we progress from lower to upper blocks, the model increasingly attends to similar regions. In the lower blocks, attention is more detailed and localized, while in the upper blocks, attention expands to cover broader regions, including those highlighted by the lower blocks. These patterns align with our intended design for visual grounding in hierarchical classification.

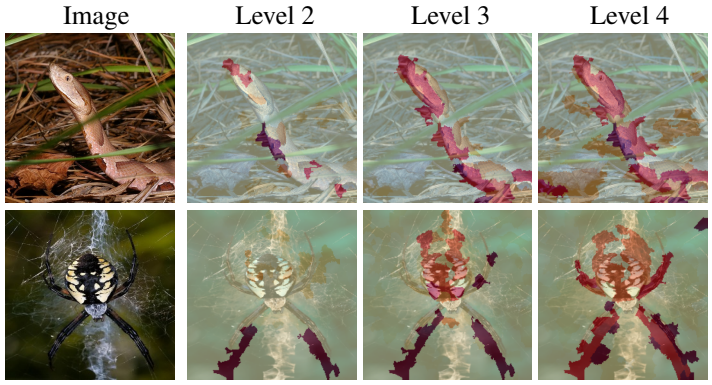


Figure 7: **Visualizations of Attention maps from H-CAST.** We align the attention weights with superpixels and average them across all heads. Darker red areas represent regions with higher attention weights. At the lower level (level 2), the attention is more focused on specific regions, such as the snake's head and parts of its body, emphasizing these as critical features for fine-grained classification (e.g., "Hypsiglena torquata"). In contrast, at the upper level (level 4), the attention expands to encompass the entire body of the snake, suggesting a shift towards a more holistic understanding of the object for coarse label (e.g., "snake"). This progression from localized to broader attention illustrates how H-CAST hierarchically integrates information across layers, supporting consistent visual grounding for hierarchical classification.

D.3 Additional Experiments for Loss Ablation

Similar to the results on the Aircraft dataset, the Living-17 dataset also shows consistent performance trends, with our proposed loss achieving strong results (Table 9, Table 10). Interestingly, for TICE, which measures only semantic consistency, the TK loss alone (Table 9) and the BCE or Flat Consistency loss achieved better performance (Table 10). However, when considering both accuracy and consistency (i.e., FPA), our proposed loss delivered the best overall performance.

Table 9: **Utilizing both losses yields best performance on Living-17.**

L_{HS}	L_{TK}	FPA	Coarse	Fine	wAP	TICE
✗	✓	84.00	90.71	84.30	86.43	**1.71**
✓	✗	84.21	90.24	84.59	86.78	2.59
✓	✓	**85.12**	**90.82**	**85.24**	**87.10**	3.19

Table 10: **KL Div. loss shows best performance on Living-17.**

Sem. Consis.	FPA	Coarse	Fine	wAP	TICE
Flat Cons.	82.82	88.88	83.53	85.31	2.51
BCE	83.65	89.76	84.00	85.92	**1.76**
KL Div.	**85.12**	**90.82**	**85.24**	**87.10**	3.19

D.4 Evaluation on CUB-200-2011 and FGVC-Aircraft datasets.

Table 11 and 12 presents results on CUB and Aircraft datasets. In our experimental results, we first observe a significant performance drop of Hier-ViT compared to Flat-ViT. This highlights a common challenge in hierarchical recognition, where training coarse and fine-grained classifiers simultaneously results in performance degradation, as observed in previous ResNet-based hierarchical recognition models (Chang et al., 2021). Our experiments reveal that this problem also exists in ViT architectures. This indicates that hierarchical recognition is a challenging problem that cannot be solely addressed by providing hierarchy supervision to class tokens. On the other hand, our method consistently outperforms most Flat models.

As Vision Transformer backbone models, when the training dataset is small, such as Aircraft and CUB with around 6K images, HRN, ResNet-based models, demonstrates better performance. However, HRN's method is highly sensitive to batch size, with a significant drop in performance observed when increasing the batch size from 8 to 64. This sensitivity makes it less suitable for training on large-scale datasets.

Nevertheless, compared to other ViT-based model, Hier-ViT, and FGN, our approach achieves significantly better performance. Specifically, using the *FPA* metric, which emphasizes both accuracy and consistency across all levels, our model achieves a remarkable improvement of +11.6%p on the Aircraft dataset and +6.3%p on the CUB dataset compared to Hier-ViT.

We also evaluate BIOCLIP Stevens et al. (2024), a foundation model for biology, on the CUB dataset, as it focuses on bird categories. BIOCLIP operates as a flat-based hierarchical model, concatenating the entire taxonomy into a single text representation. As a result, all higher-level classes are directly determined by the fine-grained species predictions, resulting in a TICE (Taxonomy-Inconsistency Error) of 0. While BIOCLIP achieves strong performance, its reliance on fine-grained predictions to define coarse classes introduces limitations in accurately predicting higher-level classes.

Table 11: **Ours consistently shows the best performance on CUB-200-2011.** It achieves 3.2 percentage point gain in FPA metric over HRN with significantly fewer parameters. Additionally, H-CAST surpasses Hier-ViT by over 6.3 percentage points. (Higher the metric is the best, except TICE.) Flat models require training 3 single models.

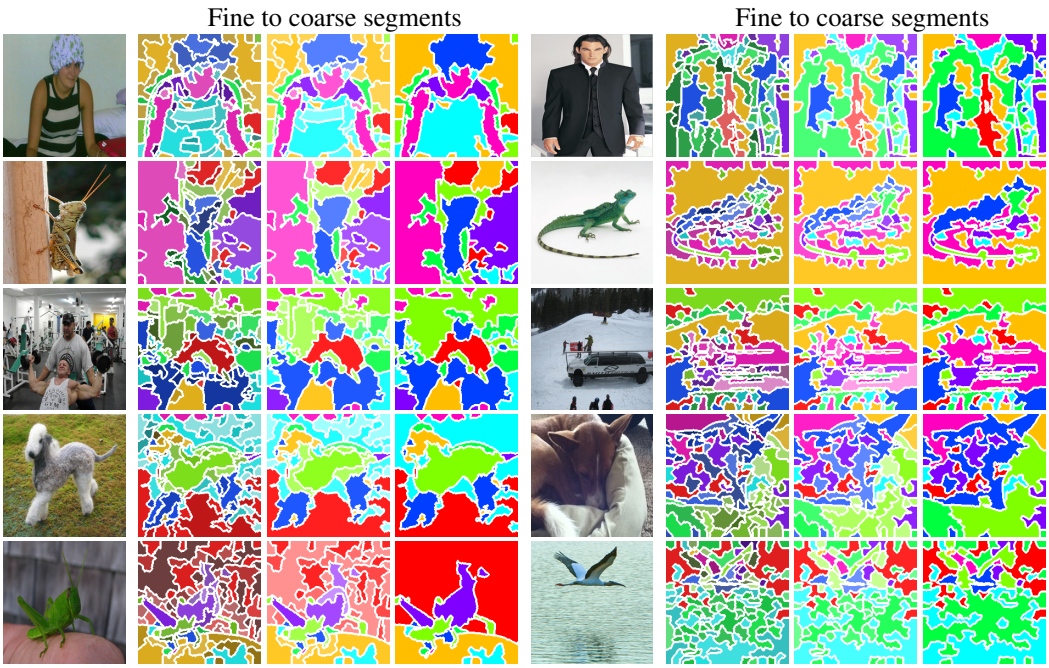
						CUB-200-2011 (13 - 38 - 200)					
		Backbone	# Params	Input image	Batch size	FPA	Order	Family	Species	wAP	TICE
Flat	Flat-ViT	ViT-S	65.1M	224x224	256	82.30	98.50	94.84	84.78	87.01	5.76
	Flat-CAST	ViT-S	78.5M	224x224	256	81.50	98.38	94.82	83.78	86.21	6.14
Hierarchy	FGN	RN-50	24.8M	224x224	128	76.08	97.05	91.44	79.29	82.05	7.73
	HRN	RN-50	94.5M	448x448	64	80.07	98.17	93.75	83.14	85.52	6.51
	HRN	RN-50	94.5M	448x448	8	**84.15**	98.58	**95.39**	**86.13**	**88.18**	4.62
	BIOCLIP (zeroshot)	ViT-B	149.6M	-	-	78.18	78.18	78.18	78.18	78.18	**0.0**
	Hier-ViT	ViT-S	21.7M	224x224	256	77.03	98.40	92.94	79.43	82.46	8.72
	Ours (H-CAST)	ViT-S	26.2M	224x224	256	83.28	**98.65**	95.12	84.86	87.13	**4.12**
	Our Gains over SOTA					-0.87	+0.07	-0.27	-1.27	-1.05	+0.50

Table 12: **Evaluation on FGVC-Aircraft.** On the smaller Aircraft dataset, ResNet-based models such as FGN and HRN show good performance. However, our H-CAST achieves better results in the consistency metric (TICE) and performs comparably in the FPA metric. Notably, H-CAST surpasses Hier-ViT by over 11 percentage points in the FPA metric.

						FGVC-Aircraft (30 - 70 - 100)					
		Backbone	# Params	Input image	Batch size	FPA	Maker	Family	Model	wAP	TICE
Flat	Flat-ViT	ViT-S	65.1M	224x224	256	76.99	94.27	91.93	80.14	86.39	10.98
	Flat-CAST	ViT-S	78.5M	224x224	256	78.22	92.95	88.93	82.39	86.26	10.77
Hierarchy	FGN	RN-50	24.8M	224x224	128	85.48	92.44	90.88	88.39	89.87	7.50
	HRN	RN-50	94.5M	448x448	64	83.56	94.93	92.68	86.59	89.97	7.26
	HRN	RN-50	94.5M	448x448	8	**91.39**	**97.15**	**95.65**	**92.32**	**94.21**	**3.36**
	Hier-ViT	ViT-S	21.7M	224x224	256	72.10	92.35	86.26	75.94	82.01	15.75
	Ours (H-CAST)	ViT-S	26.2M	224x224	256	83.72	94.96	91.39	85.33	88.90	5.01
	Our Gains over SOTA					-7.67	-2.18	-4.26	-6.99	-5.31	-1.65

D.5 Additional Visualizations of Segments

We visualize additional examples of feature grouping from fine to coarse for full-path correct and incorrect predictions on the Entity-30 dataset in Figure 8. For full-path correct predictions (all levels correct), visual details are effectively grouped to identify larger objects at coarser levels. In contrast, for full-path incorrect predictions (all levels incorrect), segments fail to recognize the object.



a) Full-Path Correct Predictions b) Full-Path Incorrect Predictions

Figure 8: Additional examples of the differences in visual grouping between cases where predictions at all levels are correct and where they are not. Correct predictions show better clustering, while incorrect predictions often exhibit fractured or misaligned groupings.

D.6 Comparison of Image Segmentation with CAST

To quantitatively evaluate the segmentation results in Figure 6, we use the ImageNet segmentation dataset, ImageNet-S (Gao et al., 2022), to obtain the ground-truth segmentation data for BREEDS dataset. The number of samples in the BREEDS validation data for which ground-truth segmentation data can be obtained from ImageNet-S is 381 for Living-17, 510 for Non-Living-26, 1,336 for Entity-30, and 1,463 for Entity-13. To calculate the region mIOU for fine-level objects, we use the last-level segments (8-way) for segmentation. Following CAST, we name the 8-way segmentations using OvSEG (Liang et al., 2023).

Also, we further visualize the segmentation results on Entity-30 in Figure 9, and show that additional taxonomy information improves segmentation. For example in the first 'bird' image, H-CAST is able to segment meaningful parts such as the face, belly, and a branch, with less fractured compared to the CAST. Thus, H-CAST delivers an improvement in segmentation with the benefits of hierarchy.

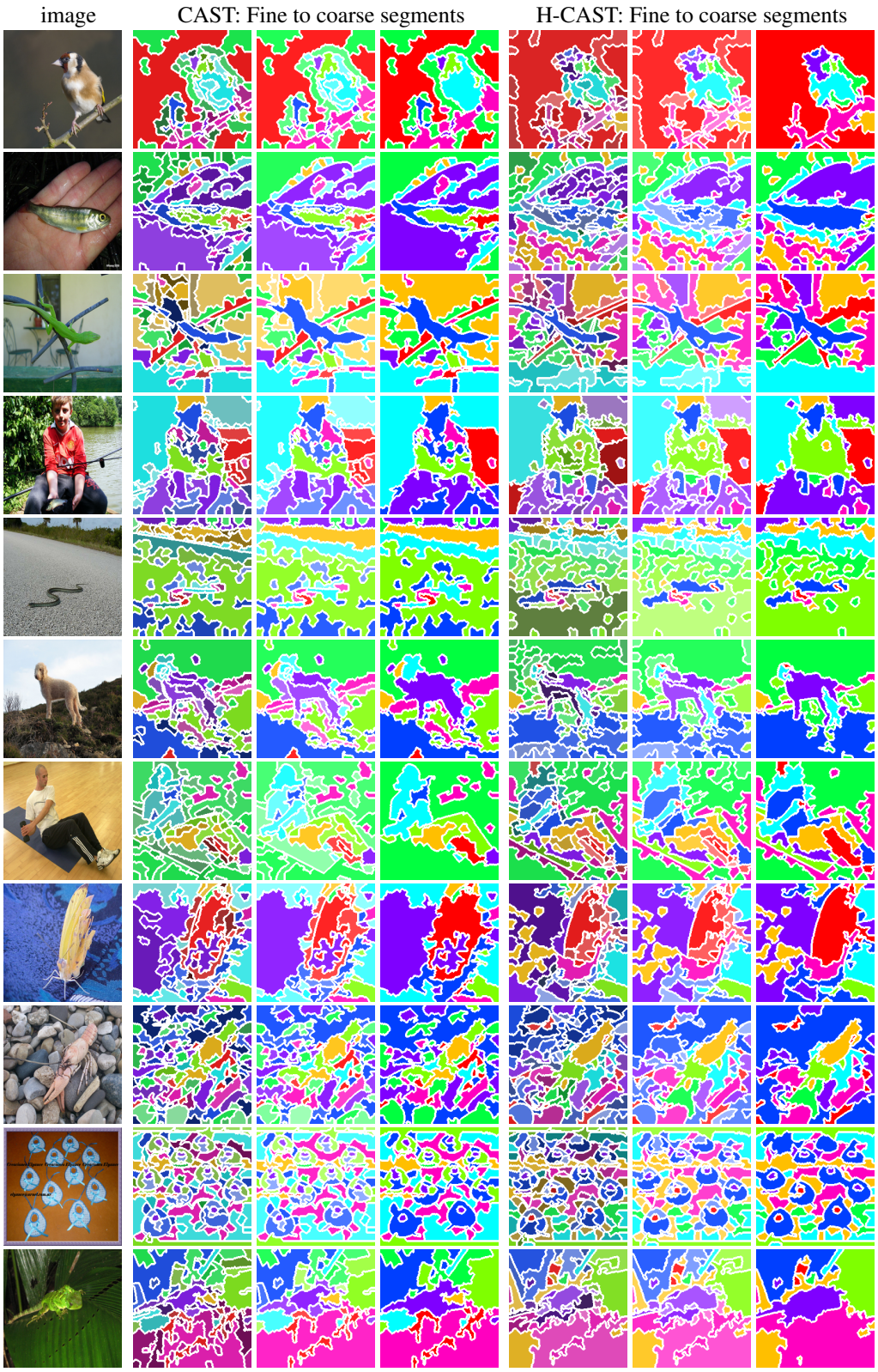


Figure 9: Additional visual results on segmentation show that H-CAST with additional taxonomy information improves segmentation. H-CAST successfully segments meaningful parts with fewer fractures compared to CAST.