Visually Consistent Hierarchical Image Classification

Seulki Park[1], Youren Zhang[1], Stella X. Yu[1,2], Sara Beery[3], Jonathan Huang[4]
[1]University of Michigan [2]UC Berkeley [3]MIT [4]Scaled Foundations
{seulki,yourenz,stellayu}@umich.edu, beery@mit.edu, jhuang11@gmail.com

Abstract

Hierarchical classification requires predicting an entire taxonomy rather than a single flat label, demanding both accurate predictions and consistency across levels. Prior methods address this by using separate classifiers for each level, but this often leads to inconsistent focus across classifiers, harming fine-grained accuracy.

We propose a novel hierarchical model that enforces pixel-level consistency by grounding fine-to-coarse semantic parsing on hierarchical visual segmentation. Instead of treating each level independently, our approach progressively groups fine-grained details (e.g., beak, wings) into coarser representations (e.g., overall bird shape), ensuring consistent feature learning. Additionally, we introduce a tree-path KL divergence loss to enforce semantic consistency. Our method achieves state-of-the-art performance on hierarchical classification benchmarks, significantly improving over zero-shot CLIP and prior approaches.

1. Introduction

Hierarchical classification aims to predict an entire taxonomy (e.g., "Birds"-"Hummingbird"-"Green Hermit") rather than a single label, making it crucial for real-world applications where the level of detail varies based on user expertise. Non-experts may require only a coarse label like "Bird," while experts need fine-grained predictions such as "Green Hermit." Flat fine-grained classification often fails in ambiguous scenarios, such as distant or motion-blurred observations, highlighting the need for **multi-granularity predictions** that improve robustness and adaptability.

A major challenge in hierarchical classification is achieving **visual consistency** across different levels of granularity. Existing methods use separate classifiers for each level, often leading to *inconsistent* predictions that do not adhere to the hierarchical taxonomy. As shown in Figure 1, separate classifiers frequently focus on different image regions, leading to misclassifications. For example, while a fine-grained classifier correctly attends to the distinguishing details of a "feather boa," the coarse classifier may incorrectly focus on irrelevant background elements, leading to a misprediction as "motor vehicle."

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

Figure 1. **Hierarchical classification suffers from inconsistent visual focus across levels, leading to misclassification.** In (a), both classifiers focus on the same object, ensuring consistent predictions. In contrast, in (b), the fine-grained classifier correctly attends to the *feather boa,* while the coarse classifier mistakenly focuses on the *bicycle*'s frame, predicting *motor vehicle*. This highlights the importance of **pixel-level visual grounding**, where classifiers at different levels align their focus while refining details. Our method enforces this visual consistency, improving both consistency in predictions and classification accuracy.

We argue that hierarchical classification can benefit from **pixel-level modeling**, ensuring that classifiers at different levels of granularity focus on the same object regions while extracting progressively finer details. Inspired by recent advances in unsupervised image segmentation, we leverage the CAST [15] method, which groups related pixels through internal parsing *without* requiring segmentation labels. Based on this, we propose **Hierarchical-CAST (H-CAST)**, which enforces **consistent visual grouping** by aligning feature extraction across hierarchical levels.

As illustrated in Figure 2, unlike prior methods [2, 12, 17, 29] that treat hierarchical classification as independent blocks, our approach integrates fine-to-coarse feature aggregation to ensure a shared visual understanding throughout the hierarchy. By progressively refining object details while maintaining a common focus, our model reduces inconsistencies and improves classification accuracy.

Additionally, we show that even vision foundation mod-

Figure 2. **While prior methods rely on separate blocks for different hierarchy levels, leading to inconsistent predictions, our integrated model enforces pixel-level consistency by progressively grouping fine-grained details into coarser representations.** 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.

els, such as CLIP [21], struggle with consistent hierarchical classification, emphasizing the need for dedicated pixel-level solutions. To evaluate this aspect, we introduce Full-Path Accuracy (FPA), a metric that measures prediction consistency across all hierarchical levels. Through extensive experiments on hierarchical benchmarks, we demonstrate that our method achieves superior consistency and accuracy compared to state-of-the-art approaches, all while requiring fewer parameters and training time. Furthermore, we show that hierarchical semantic structures can improve image segmentation, reinforcing the importance of consistent multi-scale visual modeling.

2. Consistent Hierarchical Classification

We address two key inconsistencies in hierarchical recognition: visual inconsistency, where classifiers at different levels attend to different image regions (Figure 2), and semantic inconsistency, where predictions violate the taxonomy structure (e.g., "Plant"-"Hummingbird"). To mitigate these issues, we propose H-CAST to enforce visual consistency (Section 2.1) and Tree-path KL Divergence loss to align predictions within the taxonomy (Section 2.2). An overview of our approach is provided in Figure 3.

2.1. H-CAST for Visual Consistency

The focus in an image varies between fine-grained and coarse-level classification. Fine-grained recognition (e.g., distinguishing *Green Hermit* from *Ruby-throated Hummingbird*) relies on fine details like the beak and wings, while coarse-level classification (e.g., *bird* vs. *plant*) uses broader structures like the body. However, this shift should not discard fine details but rather integrate pixel-level information across levels. Thus, hierarchical models should enforce pixel-level consistency. Based on this, we design a model that progressively refines pixel groupings, ensuring consistent visual grounding from fine to coarse levels.

We build on CAST [15], which develops a hierarchical segmentation from fine to coarse, but lacks explicit hierarchical supervision. We extend this by introducing fine-to-coarse feature grouping, where finer segments (e.g., *beak*, *wings*) contribute to coarser structures (e.g., *body*), creating a jointly structured classification and segmentation model—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 [27] 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 [15] 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

Figure 3. **Our method consists of two parts: Visual Consistency and Semantic Consistency module.** In the **Visual Consistency** module groups superpixel-parsed images from fine to coarse levels, ensuring hierarchical classifiers focus on the same regions while capturing varying granularities. The **Semantic Consistency** module incorporates label hierarchy relationships, enabling consistent learning across all levels. By promoting alignment, our method encourages classifiers at different levels to work synergistically, improving overall performance.

process. This ensures that labels from different levels progressively contribute to each other.

2.2. Tree-path KL loss for Semantic Consistency

To enhance semantic consistency, we introduce Tree-path KL Divergence loss, which directly encodes hierarchical label relationships. Instead of treating levels separately, our loss structures predictions as a hierarchical tree. We concatenate labels from all levels into a single 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. The model outputs are similarly concatenated and passed through a log softmax, and we apply KL divergence to align them with Y:

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

This loss enforces taxonomy alignment, training on all levels simultaneously. Our final loss becomes as follows, where α is a hyperparameter to control the weight of L_{TK},

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

3. Experiments

3.1. Experimental Settings

Datasets. We use widely used benchmarks in hierarchical recognition: BREEDS [23], CUB-200-2011 [32], FGVC-Aircraft [18], and iNat21-Mini [28]. Details of datasets are summarized in Appendix B.1

Evaluation Metrics. We evaluate models with following metrics, and details are in Appendix B.2.

- ***level-accuracy***: the proportion of correctly classified instances at each level [2].
- ***weighted average precision (wAP)*** [17]: adjusted accuracy by accounting for difficulty across levels.
- ***Tree-based InConsistency Error rate (TICE)*** [29]: the percentage of inconsistent paths in the taxonomy.
- ***Full-Path Accuracy (FPA)***: the proportion of samples correctly classified at all levels, serving as our primary metric for both accuracy and consistency.

Comparison methods. We evaluate H-CAST against following models (see details in Sec. B.3, Appendix): **FGN**[2], **HRN**[3], and **TransHP**[30]. We also compare a variant **Hier-ViT**, which removes visual segments and TK loss. Additionally, we benchmark against flat models like **Flat-ViT** and **Flat-CAST**[15], as well as **Hierarchical Ensembles (HiE)** [11]. For fair evaluation, we follow the official codebases and optimal settings. Further details on architectures and hyperparameters are provided in Appendix B.4.

3.2. Hierarchical Classification on Benchmarks

Table 1 shows that H-CAST outperforms all hierarchical and flat baselines, including FGN, HRN, TransHP, Hier-ViT, ViT, CAST, and HiE. Notably, H-CAST surpasses ViT-based baselines (Hier-ViT, TransHP) by over 11 percentage points, highlighting that its success comes from consistent visual grounding and Tree-path Loss, rather than simply adding hierarchy supervision to a ViT backbone. H-CAST also improves FPA by 4.3–6.4 points over HRN, despite using significantly fewer parameters. Flat models require separate training for each hierarchy level, increasing memory and training time. The strong performance of H-CAST against these baselines further underscores its effectiveness. Additional results on Aircraft, CUB (Appendix C.7), and iNaturalist 2021-Mini (Appendix C.4) show similar trends.

3.3. Enhancing Spatial Focus with H-CAST

H-CAST improves spatial consistency and interpretability by guiding feature grouping across hierarchy levels. Figure 4 (Left) shows that in full-path correct predictions, fine-level details are effectively captured and progressively grouped into larger objects at the coarse level. In contrast, in full-path incorrect predictions (Right), segments fail to form coherent objects, leading to misclassification. This highlights the benefit of incorporating segments in hierarchical classification, ensuring both consistent focus and improved interpretability. Additionally, attention map visualizations (Appendix C.2)

Table 1. **H-CAST achieves both high consistency and accuracy, outperforming all baselines on BREEDS.** It improves FPA by 4.3–6.4 points over HRN with fewer parameters and exceeds Hier-ViT and TransHP by 11+ points, highlighting the impact of consistent visual grounding and Tree-path Loss over simple hierarchy supervision in ViT. (Higher is better for all metrics except TICE.) 'ViT-S' refers to ViT-Small, and 'RN-50' denotes ResNet-50.

				Living-17 (17-34)					Non-Living-26 (26-52)				
		Backbone	# Params	FPA	Coarse	Fine	wAP	TICE	FPA	Coarse	Fine	wAP	TICE
Flat	Flat-ViT	ViT-S	65.0M	66.24	75.71	72.06	73.28	17.11	57.46	67.50	65.73	57.46	23.27
	Flat-ViT + HiE [11]	ViT-S	65.0M	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	78.82	88.06	82.88	84.61	8.82	76.17	84.77	81.08	82.31	11.77
	Flat-CAST + HiE [11]	ViT-S	78.5M	81.59	88.06	83.24	84.85	5.18	79.23	84.77	81.39	82.51	6.19
Hierarchy	FGN [2]	RN-50	24.8M	63.82	72.59	68.00	69.53	12.12	60.81	69.46	65.77	67.00	16.46
	HRN [3]	RN-50	70.8M	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	74.06	80.94	74.88	76.90	10.50	72.04	73.31	68.39	70.03	12.45
	TransHP [30]	ViT-S	21.7M	74.35	83.00	76.65	78.76	8.35	68.62	77.31	72.31	73.97	13.04
	Ours (H-CAST)	ViT-S	26.2M	**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	FPA	Coarse	Fine	wAP	TICE	FPA	Coarse	Fine	wAP	TICE
Flat	Flat-ViT	ViT-S	65.0M	64.22	76.28	76.06	76.08	21.33	66.93	76.28	74.35	74.77	18.75
	Flat-ViT + HiE [11]	ViT-S	65.0M	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	78.63	87.80	83.72	84.09	10.65	82.67	87.89	83.15	84.73	5.26
	Flat-CAST + HiE [11]	ViT-S	78.5M	79.52	87.80	83.40	83.80	6.83	83.70	87.89	84.30	85.02	4.20
Hierarchy	FGN [2]	RN-50	24.8M	74.23	85.35	78.00	78.67	9.43	68.52	77.47	73.18	74.04	13.62
	HRN [3]	RN-50	70.8M	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	74.63	86.95	75.39	77.70	5.19	73.01	81.38	74.10	74.76	11.61
	TransHP [30]	ViT-S	21.7M	73.45	86.28	76.23	77.14	8.80	72.00	80.78	75.63	76.66	12.20
	Ours (H-CAST)	ViT-S	26.2M	**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

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

Figure 4. **H-CAST enhances the interpretability of the model's predictions.** We visualize how fine-to-coarse segmentations evolve for correct and incorrect full-path predictions on Entity-30. When predictions are correct, visual details are coherently grouped, forming meaningful objects at coarser levels. In contrast, incorrect predictions exhibit fragmented segments, failing to capture object structure. For instance, in the first row, the correctly classified shoe is grouped with the ankle in green, while the misclassified case shows disjointed segments. This highlights the role of visual segmentsin guiding focus and enhancing interpretability in hierarchical recognition.

show that lower blocks attend to fine details while upper blocks capture broader structures, reinforcing consistent visual grounding. We include further ablation studies and analysis in Appendix.

4. Conclusion

We tackle hierarchical classification inconsistencies through pixel-level visual grounding, ensuring classifiers focus on coherent regions across levels. Our approach improves both accuracy and consistency, as demonstrated on benchmark datasets.

References

[1] Luca Bertinetto, Romain Mueller, Konstantinos Tertikas, Sina Samangooei, and Nicholas A Lord. Making better mistakes: Leveraging class hierarchies with deep networks. In *Proceed-*

ings of the IEEE/CVF conference on computer vision and pattern recognition, 2020. 7

[2] 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. 1, 3, 4, 7, 8, 9, 16

[3] 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. 3, 4, 7, 9

[4] 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. 9

[5] 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. 7

[6] Shanghua Gao, Zhong-Yu Li, Ming-Hsuan Yang, Ming-Ming Cheng, Junwei Han, and Philip Torr. Large-scale unsupervised semantic segmentation. *TPAMI*, 2022. 12

[7] Ashima Garg, Depanshu Sani, and Saket Anand. Learning hierarchy aware features for reducing mistake severity. In *European Conference on Computer Vision*, 2022. 7

[8] Vivien Sainte Fare Garnot and Loic Landrieu. Leveraging class hierarchies with metric-guided prototype learning. *arXiv preprint arXiv:2007.03047*, 2020. 14

[9] Grant Van Horn, Oisin Mac Aodha, Yang Song, Yin Cui, Chen Sun, Alexander Shepard, Hartwig Adam, Pietro Perona, and Serge J. Belongie. The inaturalist species classification and detection dataset. In *2018 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2018, Salt Lake City, UT, USA, June 18-22, 2018*, 2018. 14

[10] 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. 7

[11] 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. 3, 4, 9

[12] Juan Jiang, Jingmin Yang, Wenjie Zhang, and Hongbin Zhang. Hierarchical multi-granularity classification based on bidirectional knowledge transfer. *Multimedia Systems*, 2024. 1, 7

[13] 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. 7

[14] 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. 7

[15] 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. 1, 2, 3, 7, 9

[16] 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. 12

[17] 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. 1, 3, 7, 8

[18] S. Maji, J. Kannala, E. Rahtu, M. Blaschko, and A. Vedaldi. Fine-grained visual classification of aircraft. Technical report, 2013. 3, 8

[19] George A Miller. Wordnet: a lexical database for english. *Communications of the ACM*, 1995. 8

[20] 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. 7

[21] 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, 10

[22] 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. 8

[23] Shibani Santurkar, Dimitris Tsipras, and Aleksander Madry. {BREEDS}: Benchmarks for subpopulation shift. In *International Conference on Learning Representations*, 2021. 3, 8

[24] 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. 7, 16

[25] 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. 9

[26] Hugo Touvron, Matthieu Cord, Matthijs Douze, Francisco Massa, Alexandre Sablayrolles, and Herve Jegou. Training

data-efficient image transformers & distillation through attention. In *Proceedings of the 38th International Conference on Machine Learning*, 2021. 9

[27] 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. 2

[28] 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. 3, 8, 14

[29] 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, 7, 8

[30] Wenhao Wang, Yifan Sun, Wei Li, and Yi Yang. TransHP: Image classification with hierarchical prompting. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023. 3, 4, 7, 9, 14, 15

[31] Jonatas Wehrmann, Ricardo Cerri, and Rodrigo Barros. Hierarchical multi-label classification networks. In *International conference on machine learning*. PMLR, 2018. 7

[32] 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. 3, 8

[33] Zhenlin Xu, Yi Zhu, Siqi Deng, Abhay Mittal, Yanbei Chen, Manchen Wang, Paolo Favaro, Joseph Tighe, and Davide Modolo. Benchmarking zero-shot recognition with vision-language models: Challenges on granularity and specificity. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024. 10

[34] 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. 15

[35] 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. 9

[36] 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. 7

[37] Hongyi Zhang, Moustapha Cisse, Yann N Dauphin, and David Lopez-Paz. mixup: Beyond empirical risk minimization. *arXiv preprint arXiv:1710.09412*, 2017. 9

[38] 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. 7, 14

[39] Xinqi Zhu and Michael Bain. B-cnn: branch convolutional neural network for hierarchical classification. *arXiv preprint arXiv:1709.09890*, 2017. 7, 15

Supplementary Material

A. Related Work

Hierarchical classification presents a unique challenge: the input image remains the same, but the output varies across a **semantic hierarchy** (e.g., "Bird" → "Hummingbird" → "Green Hermit"). Prior research has predominantly focused on encoding this hierarchy in the semantic (text) space, employing methods such as hierarchical loss functions [1, 36] or modeling taxonomies as flat, structured text inputs, as seen in BIOCLIP [24]. Existing approaches can be broadly categorized into two groups:

1. Flat-level output methods, which predict only the most fine-grained label and infer coarse categories afterward [1, 5, 7, 13, 30, 36, 38]. These methods struggle in ambiguous test-time scenarios where fine-grained details are not clearly visible, making fine-grained classification fundamentally infeasible.

2. Full-hierarchy output methods, which predict an entire taxonomy in a single forward pass [2, 3, 12, 17, 31, 39]. The most common strategy uses a shared backbone with separate branches per hierarchy level. While effective, these independent branches can lead to inconsistent predictions across levels, as they process images separately rather than integrating multi-level visual cues [29]. In contrast to these approaches, we take a visual-space perspective and investigate how hierarchical classification can be improved by enforcing pixel-level consistency across classification levels. To the best of our knowledge, no prior work has explored visual grounding via unsupervised segmentation to enhance consistency in hierarchical classification.

Unsupevised/Weakly-supervised Semantic Segmentation aims to group pixels without pixel-level annotations or using only class labels [10, 14, 15, 20]. These methods employ hierarchical grouping, where smaller parts (e.g., a face or arm) merge into larger structures (e.g., a whole body).

Building on this, we leverage CAST [15], which naturally supports consistent visual grouping. Unlike prior work, we introduce the novel insight that spatial granularity in segmentation aligns with taxonomy hierarchies, where fine-grained classifiers benefit from finer segments, while coarse classifiers utilize broader groupings—a connection not previously explored.

B. Experimental Settings

B.1. Benchmark Datasets

We use widely used benchmarks in hierarchical recognition: BREEDS [23], CUB-200-2011 [32], FGVC-Aircraft [18], and iNat21-Mini [28]. BREEDS, a subset of ImageNet [22], includes four 2-level hierarchy datasets with different depths/parts based on the WordNet [19] 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); For experiments on a larger dataset, we used the 3-level iNat21-Mini. Table 2 summarizes a description of the datasets. Details of the iNat21-Mini are provided in Sec. C.4.

Table 2. **Benchmark Datasets.**

Datasets	L-17	NL-26	E-13	E-30	CUB	Aircraft	iNat21-Mini
# Levels	2	2	2	2	3	3	3
# of classes	17-34	26-52	13-130	30-120	13-38-200	30-70-100	273-1,103-10,000
# Train images	44.2K	65.7K	167K	154K	5,994	6,667	500K
# Test images	1.7K	2.6K	6.5K	6K	5,794	3,333	100K

B.2. Evaluation Metrics

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

- ***level-accuracy***: the proportion of correctly classified instances at each level [2].
- ***weighted average precision (wAP)*** [17]: 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)*** [29]: 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)***: 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 3.

Table 3. **FPA considers both correctness and consistency.** While TICE [29] measures only consistency, FPA marks predictions as positive only when they are both correct and consistent.

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

B.3. Comparison methods

First, we evaluate our H-CAST with representative models in hierarchical classification, **FGN** [2] and **HRN** [3]. **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. We also compare **TransHP** [30], a ViT-based model that learns prompt tokens to represent coarse classes and injects them into an intermediate block to enhance fine-grained predictions. Lastly, we compare **Hier-ViT**, a variant without visual segments and TK loss. Like 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 [15]. We also compare with **Hierarchical Ensembles, HiE** [11], 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. For baseline methods, we use the official codebases and their reported optimal settings. All models are trained for 100 epochs, except TransHP, which is trained for 300 epochs as in the original paper.

B.4. Hyperparameters for Training.

For a fair comparison, we use ViT-Small and CAST-Small models of corresponding sizes. As in CAST, we train both ViT and CAST using DeiT framework [26], and 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. Following the literature [3], we use ImageNet pre-trained models for the Aircraft, CUB, and iNat datasets. For the ImageNet subset BREEDS dataset, we train the models from scratch. We show hyper-parameter settings in Table 4.

Table 4. **Hyper-parameters for training H-CAST and ViT on FGVC-Aircraft, CUB-200-2011, BREEDS, and iNaturalist datasets.** We follow mostly the same set up as CAST [15].

Parameter	Aircraft	CUB, BREEDS, iNaturalist
batch_size	256	256
crop_size	224	224
learning_rate	$1e^{-3}$	5^{e-4}
weight_decay	0.05	0.05
momentum	0.9	0.9
total_epochs	100	100
warmup_epochs	5	5
warmup_learning_rate	$1e^{-4}$	$1e^{-6}$
optimizer	Adam	Adam
learning_rate_policy	Cosine decay	Cosine decay
augmentation [4]	RandAug(9, 0.5)	RandAug(9, 0.5)
label_smoothing [25]	0.1	0.1
mixup [37]	0.8	0.8
cutmix [35]	1.0	1.0
α (weight for TK loss)	0.5	0.5
ViT-S: # Tokens	$[196]_{\times 11}$	
CAST-S: # Tokens	$[196]_{\times 3}, [64]_{\times 3}, [32]_{\times 3}, [16]_{\times 2}$	

C. Analysis and additional studies

C.1. Hierarchical Classification with Vision Foundation Models

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

Figure 5. **Consistent predictions for hierarchical classification cannot be easily solved using a vision foundation model.** We evaluate the performance of CLIP [21] 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.

First, to demonstrate that longstanding hierarchical classification is not easily solved by today's vision foundation models, we evaluate CLIP [21]'s performance on the 2-level BREEDS dataset. The top row of Figure 5 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 5 (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 5 (c) further highlights significant errors in coarse predictions when addressing broader concepts. Our findings support the recent study from [33] that VLMs excel at fine-grained prediction but struggle with general concepts. This highlights the ongoing challenges of hierarchical classification, even with vision foundation models. Furthermore, when we examine the misclassification examples in bottom row of Figure 5, 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.

C.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 6. 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 6. **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.

C.3. 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. To quantitatively evaluate the segmentation results in Figure 7, we use the ImageNet segmentation dataset, ImageNet-S [6], 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 [16].

Figure and Table 7 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.

Also, we further visualize the segmentation results on Entity-30 in Figure 8, 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.

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 7. **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 [6]. 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.

Figure 8. 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.

C.4. Evaluation on the large-scale iNaturalist datasets.

We present the results of our experiments on the large-scale dataset, iNaturalist21-mini [28] and iNaturalist-2018 [9]. First, iNaturalist21-mini contains 10,000 classes, 500,000 training samples, and 100,000 test samples, organized within an 8-level hierarchy. For our experiments, we focused on a 3-level hierarchy consisting of *name*, *family*, and *order*. This selection was motivated by the need for models with practical and meaningful granularity for real-world applications.

The number of classes at each hierarchical level is as follows: Kingdom (3), Supercategory (11), Phylum (13), Class (51), Order (273), Family (1,103), Genus (4,884), Name (10,000)

We excluded coarse-grained levels such as *kingdom* (3 classes), *Supercategory* (11 classes), because their minimal granularity adds little value for classification tasks. Similarly, overly fine-grained levels such as *genus* (4,884 classes), where many species are represented by only one or two samples, offer limited differentiation from direct *name*-level classification. Instead, we focused on *order* (273 classes), *family* (1,103 classes), and *name* (10,000 classes) to ensure that each higher-level class represents a meaningful subset of lower-level classes, allowing for interpretable and consistent predictions.

The results are shown in Table 5. Compared to Hier-ViT, which uses the same ViT-small backbone, our method achieves a 7.29% improvement in fine-grained accuracy and an 8.3% improvement in the FPA metric, demonstrating a significant performance advantage.

Table 5. **Our H-CAST outperforms ViT-backbone baselines, Hier-ViT and TransHP, on the large-scale iNaturalist2021-Mini dataset, achieving significantly higher accuracy and consistency.**

	iNaturalist2021-Mini (273 - 1,103 - 10,000)					
	FPA	Order	Family	Name	wAP	TICE
Hier-ViT	56.73	87.54	79.79	62.81	65.05	24.34
TransHP	58.94	83.49	82.15	68.82	70.46	24.63
Ours (H-CAST)	**65.03**	**89.84**	**84.12**	**70.09**	**71.92**	**15.92**
Our Gains	+6.09	+2.30	+1.97	+1.27	+6.88	+8.42

We further compare our method with single-level approaches that utilize hierarchical labels to enhance fine-grained accuracy (*e.g.*, Guided [8], HiMulConE [38], and TransHP [30]). The comparison is conducted on the large-scale iNaturalist-2018 dataset, following TransHP. iNaturalist-2018 includes two-level hierarchical annotations with 14 super-categories and 8,142 species, comprising 437,513 training images and 24,426 validation images. We use the same H-CAST-small and the model is trained for 100 epochs, using the same hyper-parameters in Table 4. As shown in Table 6, our method achieves strong fine-grained accuracy, demonstrating the effectiveness of consistent visual grounding.

Table 6. **H-CAST outperforms methods leveraging hierarchical labels for fine-grained accuracy on the large-scale iNaturalist-2018 dataset, demonstrating the effectiveness of visual consistency.** The results are reported from TransHP [30].

	iNaturalist-2018 (Acc.)
Guided	63.11
HiMulConE	63.46
TransHP	64.21
H-CAST	**67.13**

C.5. Architecture Design

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 prior methods, which typically learn coarse features first [30, 34, 39]. 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 7 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 7. **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

C.6. 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 10 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 to match them with the ground truth. As shown in Table 11 on Aircraft dataset, KL Divergence loss achieves the best FPA, demonstrating superior accuracy and consistency.

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 8, Table 9). Interestingly, for TICE, which measures only semantic consistency, the TK loss alone (Table 8) and the BCE or Flat Consistency loss achieved better performance (Table 9). However, when considering both accuracy and consistency (i.e., FPA), our proposed loss delivered the best overall performance.

Table 8. **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 9. **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

Table 10. **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 11. **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**

C.7. Evaluation on CUB-200-2011 and FGVC-Aircraft datasets.

Table 12 and 13 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 [2]. 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.

Compared to ViT-backbone models, Hier-ViT and TransHP, our approach achieves significantly better performance. Specifically, using the *FPA* metric, which captures both accuracy and consistency across all levels, our model outperforms TransHP by +8.2%p on the Aircraft dataset and +2.6%p on the CUB dataset.

We also evaluate BIOCLIP [24], 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.

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.

Table 12. **Ours consistently shows the best performance on CUB-200-2011.** H-CAST outperforms ViT-backbone models, Hier-ViT and TransHP, by over 6.3 and 2.6 percentage points, respectively. Additionally, it achieves a 3.2 percentage point gain in the FPA metric over the ResNet-based HRN while using significantly fewer parameters. (A higher metric indicates better performance, except for TICE.) Flat models require training three separate 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**	<u>98.58</u>	**95.39**	**86.13**	**88.18**	<u>4.62</u>
	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
	TransHP	ViT-S	21.7M	224x224	128	80.70	96.70	94.15	84.59	86.66	7.16
	Ours (H-CAST)	ViT-S	26.2M	224x224	256	<u>83.28</u>	**98.65**	<u>95.12</u>	<u>84.86</u>	<u>87.13</u>	**4.12**
	Our Gains over SOTA					-0.87	+0.07	-0.27	-1.27	-1.05	+0.50

Table 13. **Evaluation on FGVC-Aircraft.** Notably, H-CAST outperforms ViT-backbone models, Hier-ViT and TransHP, by over 11 and 8 percentage points, respectively, in the FPA metric. 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.

						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	<u>85.48</u>	92.44	90.88	<u>88.39</u>	89.87	7.50
	HRN	RN-50	94.5M	448x448	64	83.56	94.93	<u>92.68</u>	86.59	<u>89.97</u>	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
	TransHP	ViT-S	21.7M	224x224	128	75.49	90.16	87.46	81.46	84.86	13.95
	Ours (H-CAST)	ViT-S	26.2M	224x224	256	83.72	<u>94.96</u>	91.39	85.33	88.90	<u>5.01</u>
	Our Gains over SOTA					-7.67	-2.18	-4.26	-6.99	-5.31	-1.65

C.8. 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 9. 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 9. 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.