ICLR2025: VISUALLY CONSISTENT HIERARCHICAL IMAGE CLASSIFICATION

J. Yang[1]

[1]Tianajin University

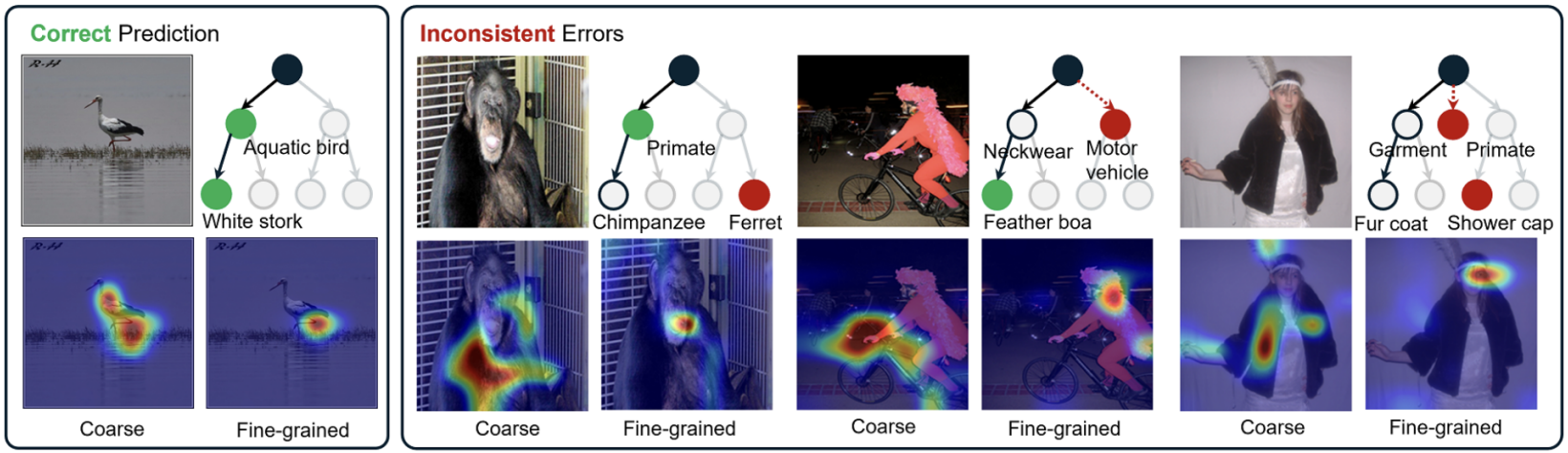
Seminar, Nov 2025

Table of Contents

Table of Contents

Inconsistent Visual Focus Hurts Hierarchical Accuracy

- Prior methods: external semantic constraints, e.g., label relation graphs
- They don't enforce visual consistency at test time
- Result: classifiers at different levels attend to unrelated regions



a) Coarse ✓, Fine ✓ **b)** Coarse ✓, Fine ✗ **c)** Coarse ✗ , Fine ✓ **d)** Coarse ✗, Fine ✗

Conclusion

Attention is all you need.

Insight: Visual Consistency Enables Coherent Prediction

- Align attention across coarse-to-fine levels
- Use intra-image segmentation to enforce visual grounding
- Shared visual parsing avoids contradiction across levels

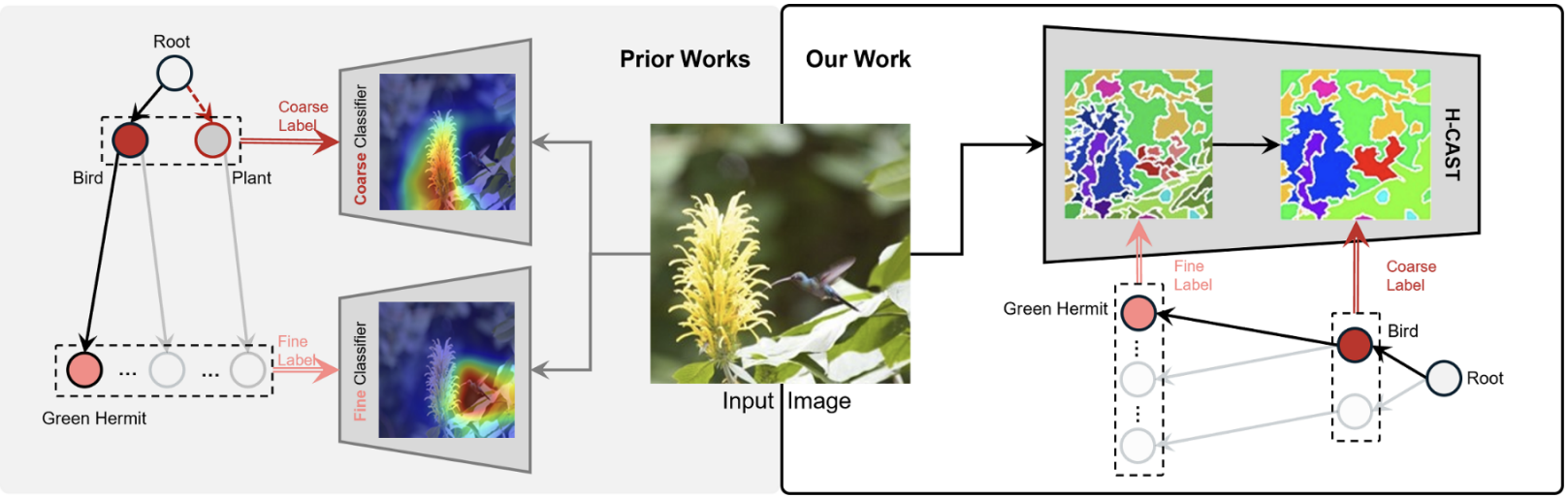


Figure: Consistent Focus

Method Overview: H-CAST Architecture

- Builds on CAST (hierarchical segmentation backbone)
- Adds classification heads at different segmentation levels
- Supervision is applied progressively: fine to coarse

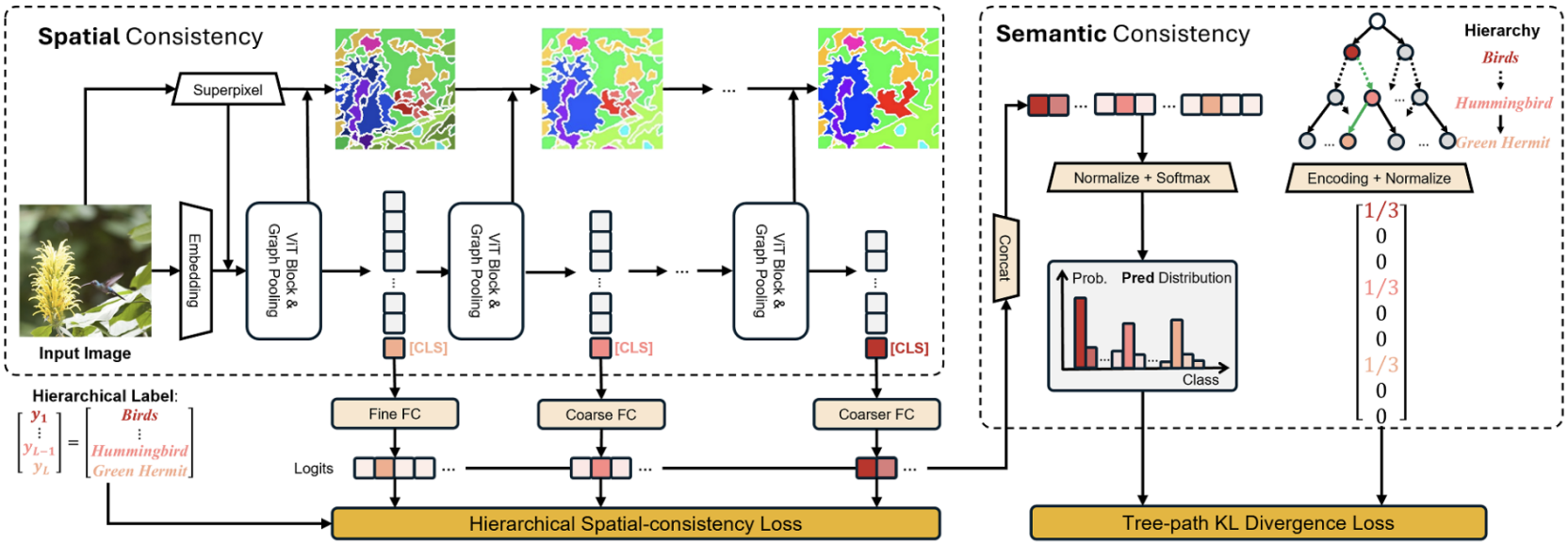


Figure: Model Architecture

Key Components of H-CAST

1. **Visual Consistency Module**
   - Segment image into fine-to-coarse hierarchy
   - Enforce consistent attention via classification heads at each level
2. **Semantic Consistency Module**
   - Tree-path KL Divergence Loss
   - Encourages label path alignment: parent-child compatibility

The Novel Module: Graph Pooling

THE KEY THOUGHT IS CLUSTERING!

- Initialize several cluster center by Farthest Point Sampling(FPS)
- Use Cosine Similarity to calculate the similarity on features
- Change the similarity to assignment matrix by softmax

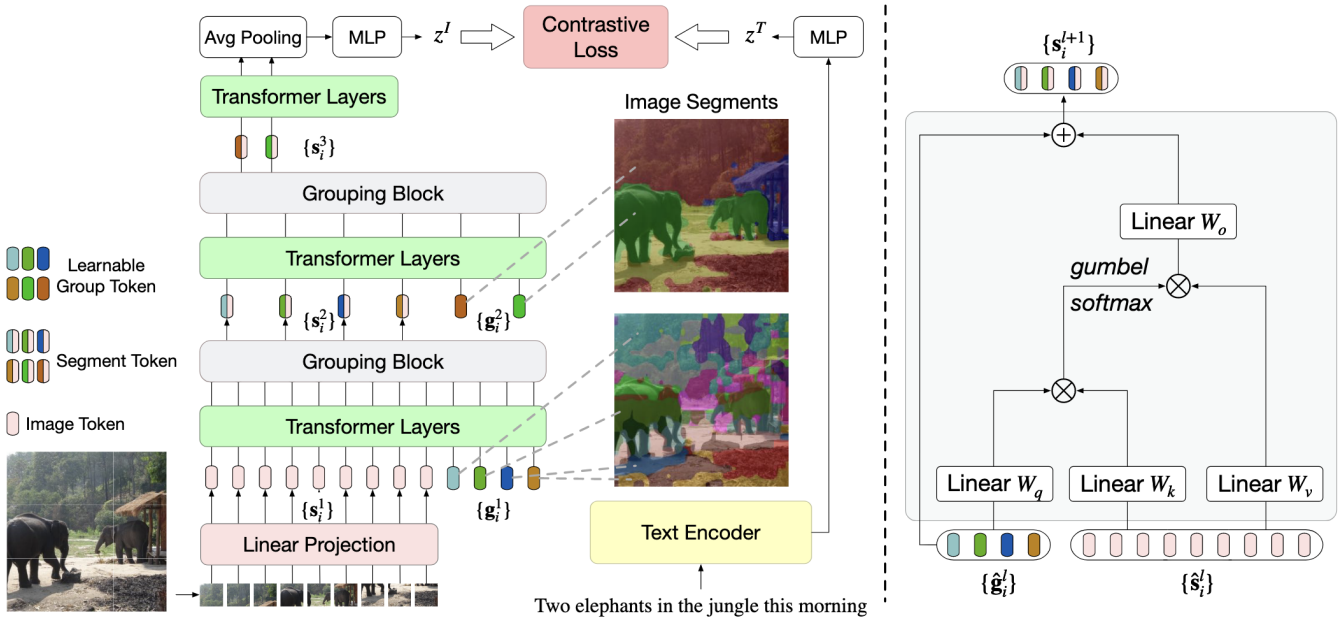
$$Z_l \Leftarrow Z_l^{init} + \mathrm{MLP}\left(P_l^\top Z_{l-1} ./ P_l^\top \mathbf{1}\right) \tag{1}$$

where Z_l is the feature for current layer, Z_l^{init} is the feature selected by FPS from former layer, Z_{l-1} is the feature from the former layer, P_l^T is the assignment matrix

The Novel Module: Graph Pooling

Advantages compared to GroupViT

- The # of Cluster can be changed in inference period without fine-tuning.
- Super-pixel is much finer compared to the square patches, which makes the segmentation more precise.
- No Gradient Approximation

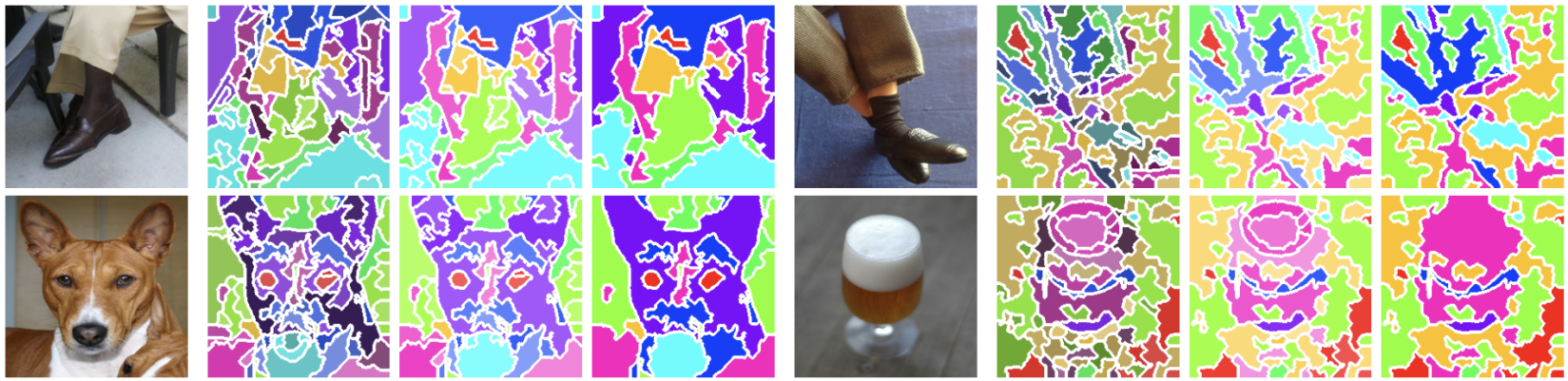


(a) GroupViT Architecture and Training Pipeline

(b) Grouping Block

Why Structured Visual Parsing Helps

- Successful predictions show coherent object groupings
- Failed predictions show fragmented, incoherent segments

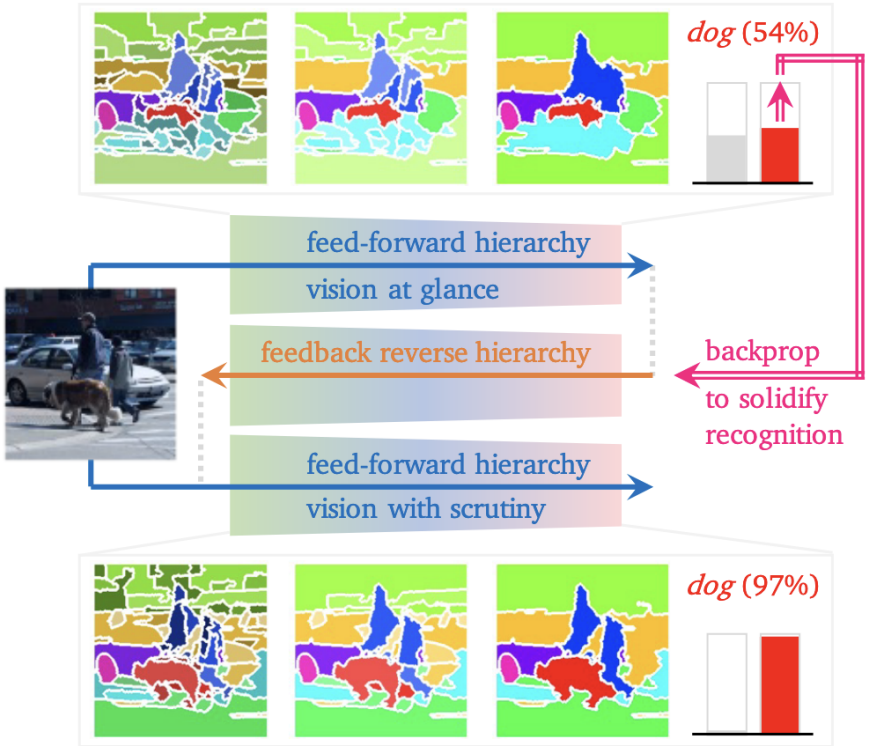


(a) Full-Path Correct Predictions

(b) Full-Path Incorrect Predictions

Case Optimization

With targeted feedback back-propagating in a reverse hierarchy, it refines internal part-to-whole segmentation by recognition.



After back-propagating to increase dog activation, the model undergoes test-time adaptation in a reverse hierarchy. This adjustment allows the next feed-forward process to uncover the whole dog and boost dog activation to 97%!