

Visually Consistent Hierarchical Image Classification



Seulki Park



Youren Zhang



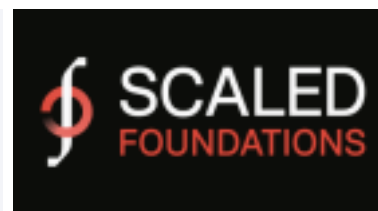
Stella X. Yu



Sara Beery



Jonathan Huang



Fine-grained Classification at a Single Level



...



...



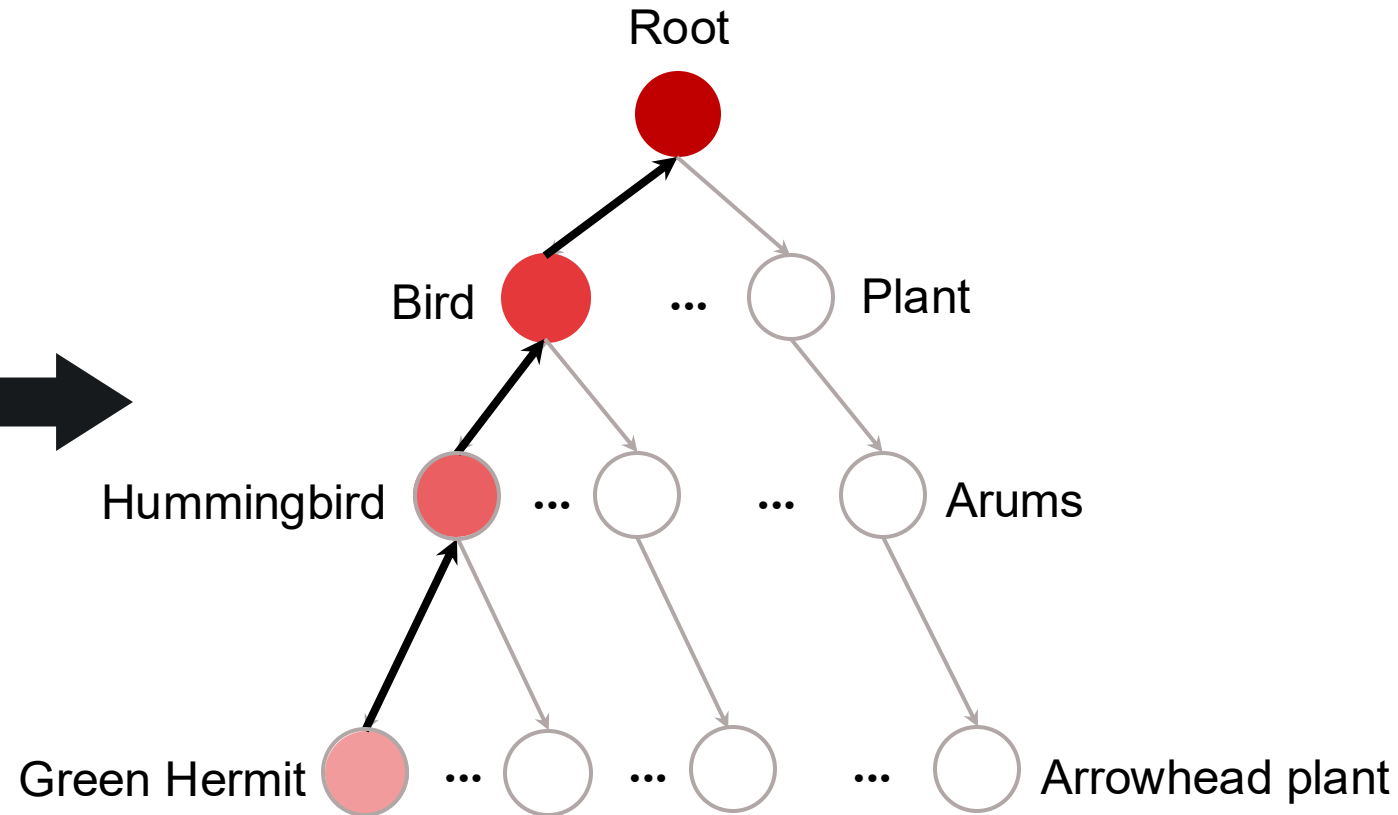
...



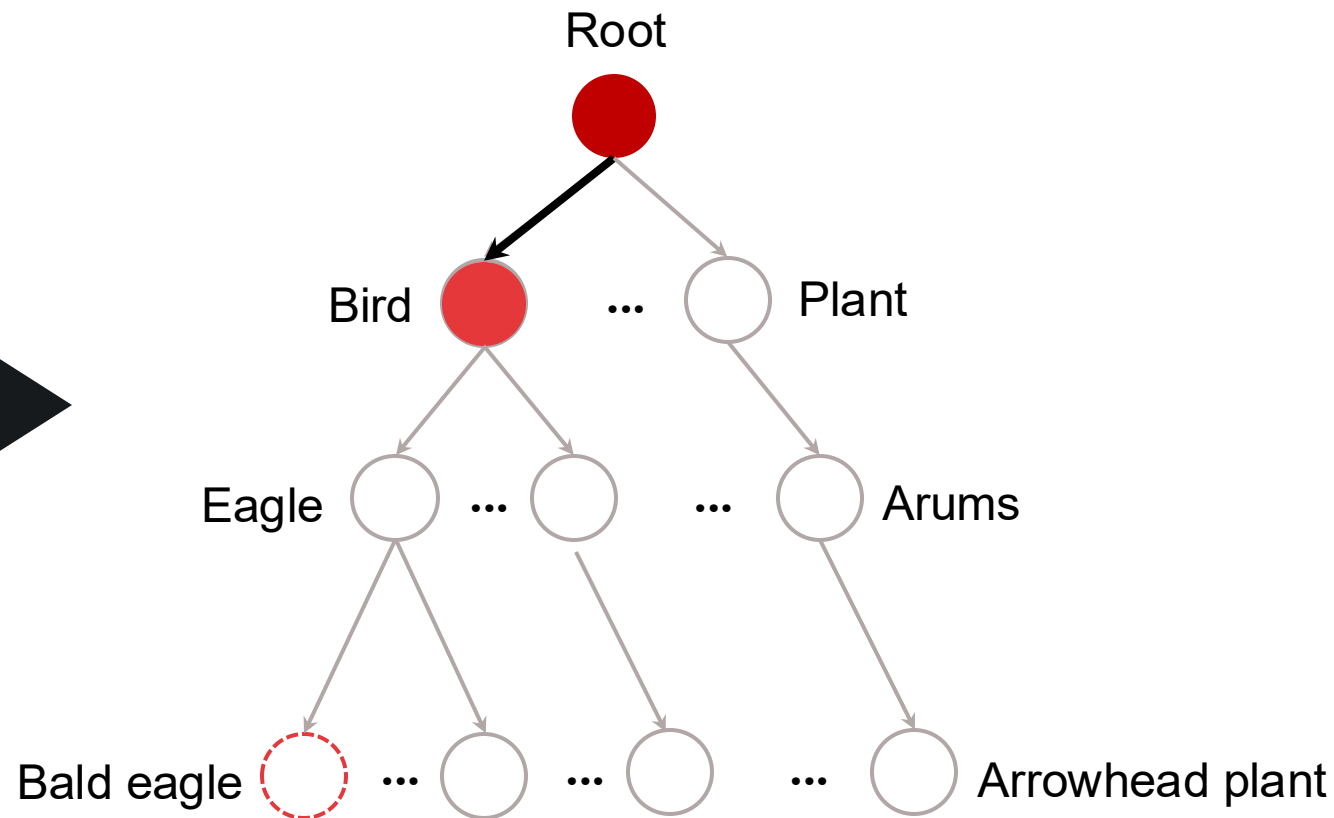
Green Hermit

Arrowhead plant

Hierarchical Image Classification

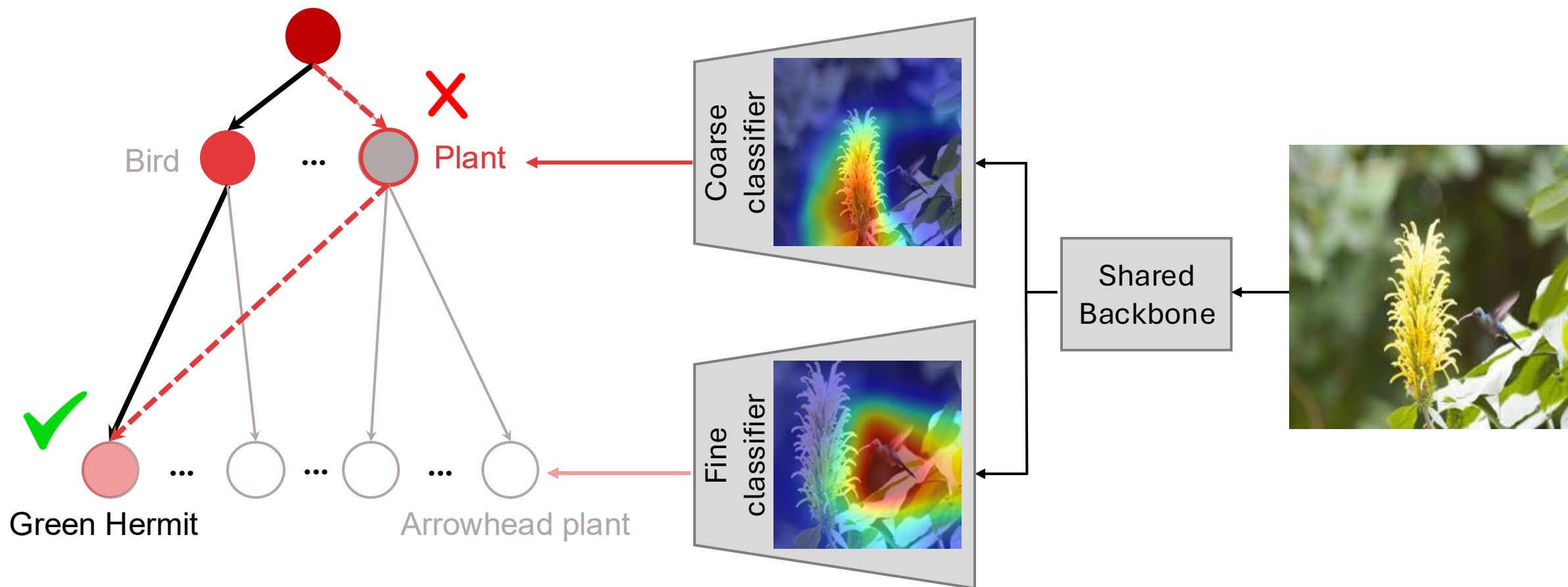


What if Fine-grained Prediction is Infeasible?

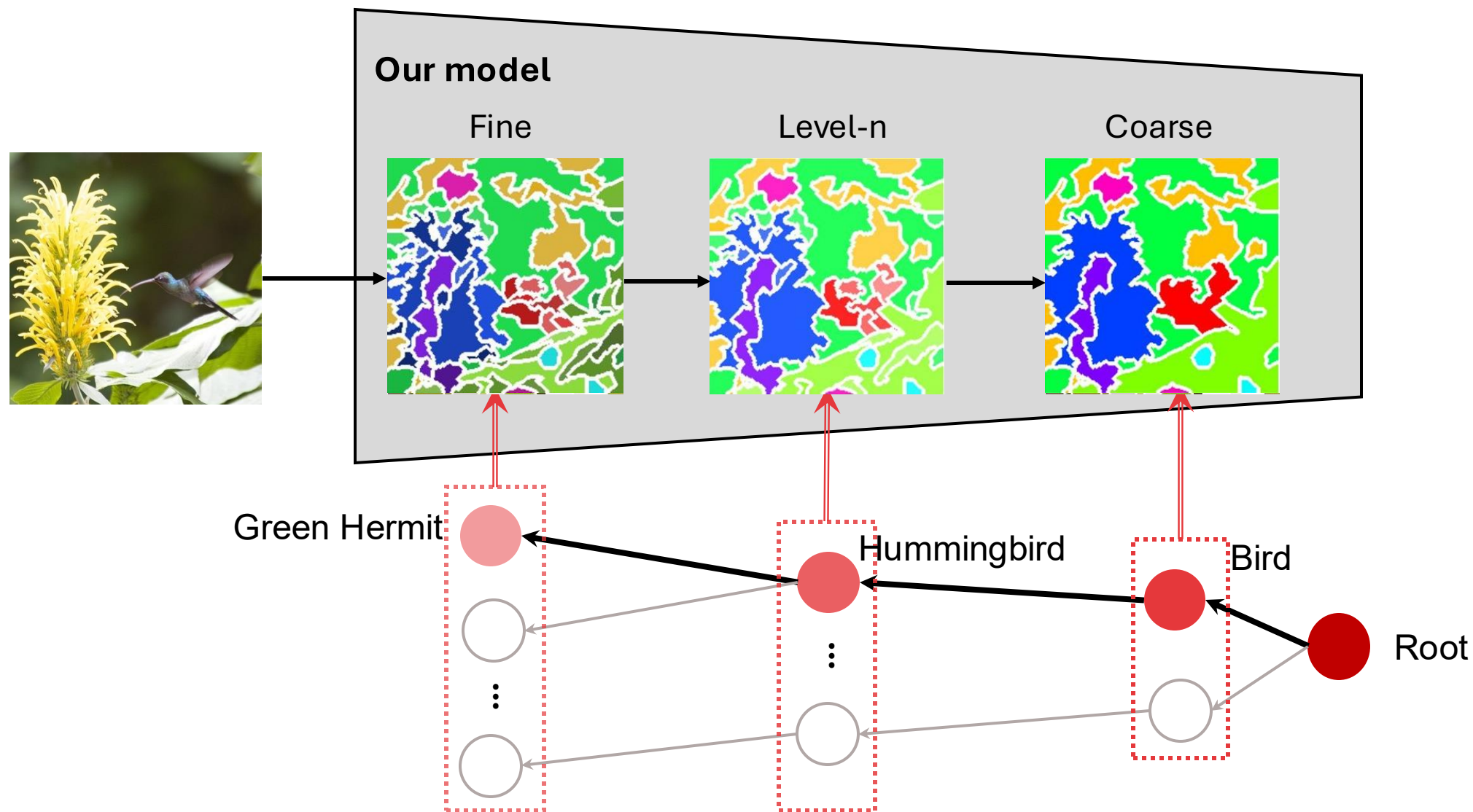


Impossible to make a fine-grained prediction!

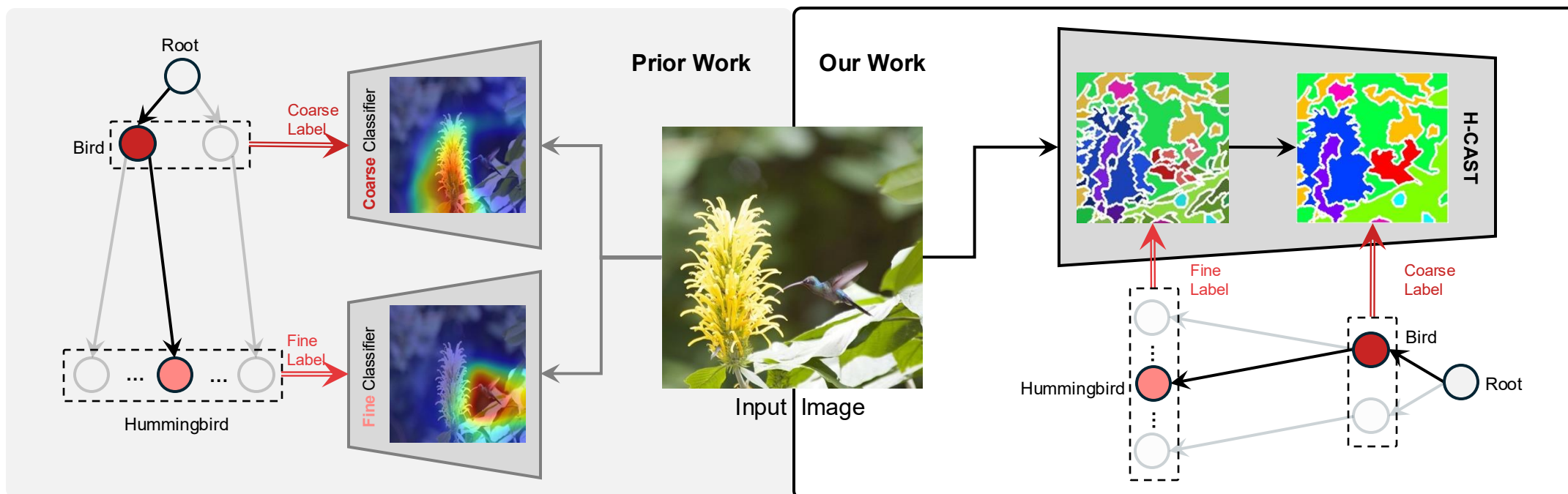
Predicting Each Level Leads to a Multi-task Problem



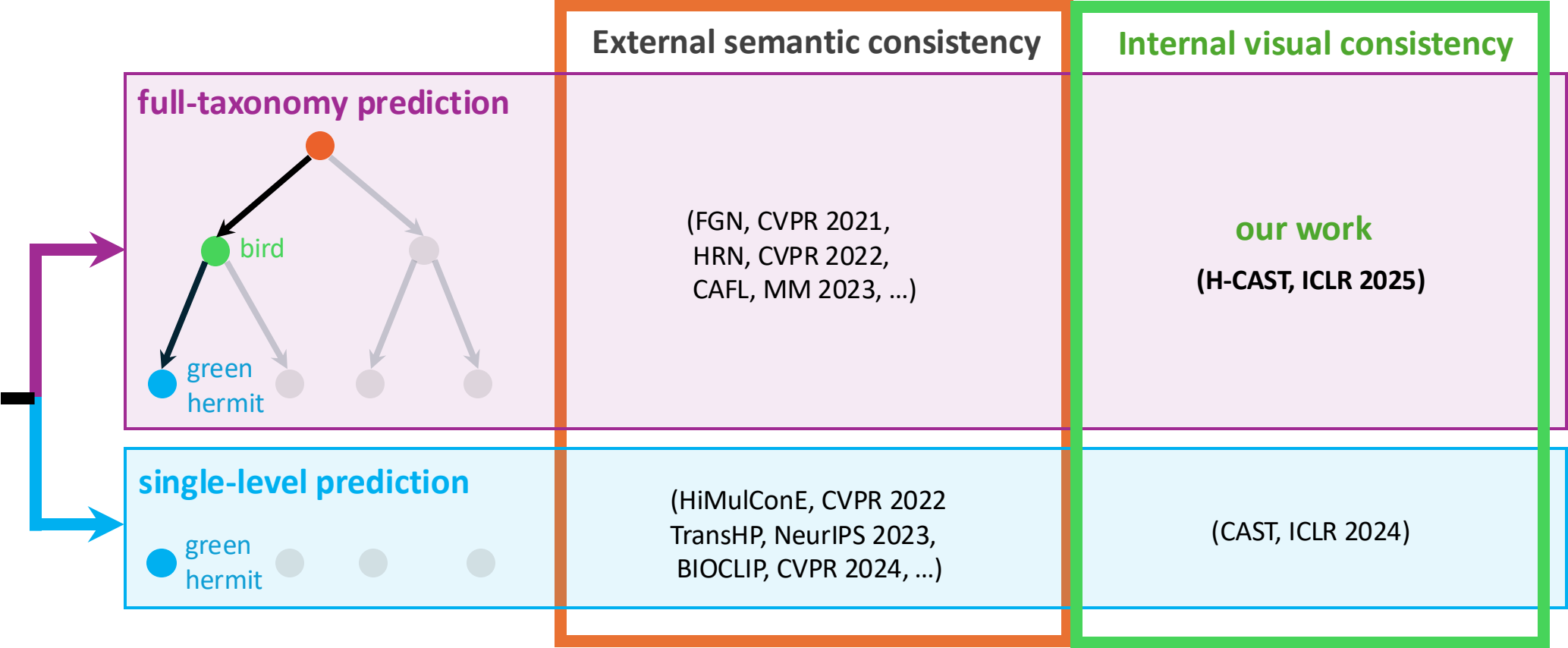
Our Idea: Internal Visual Consistency



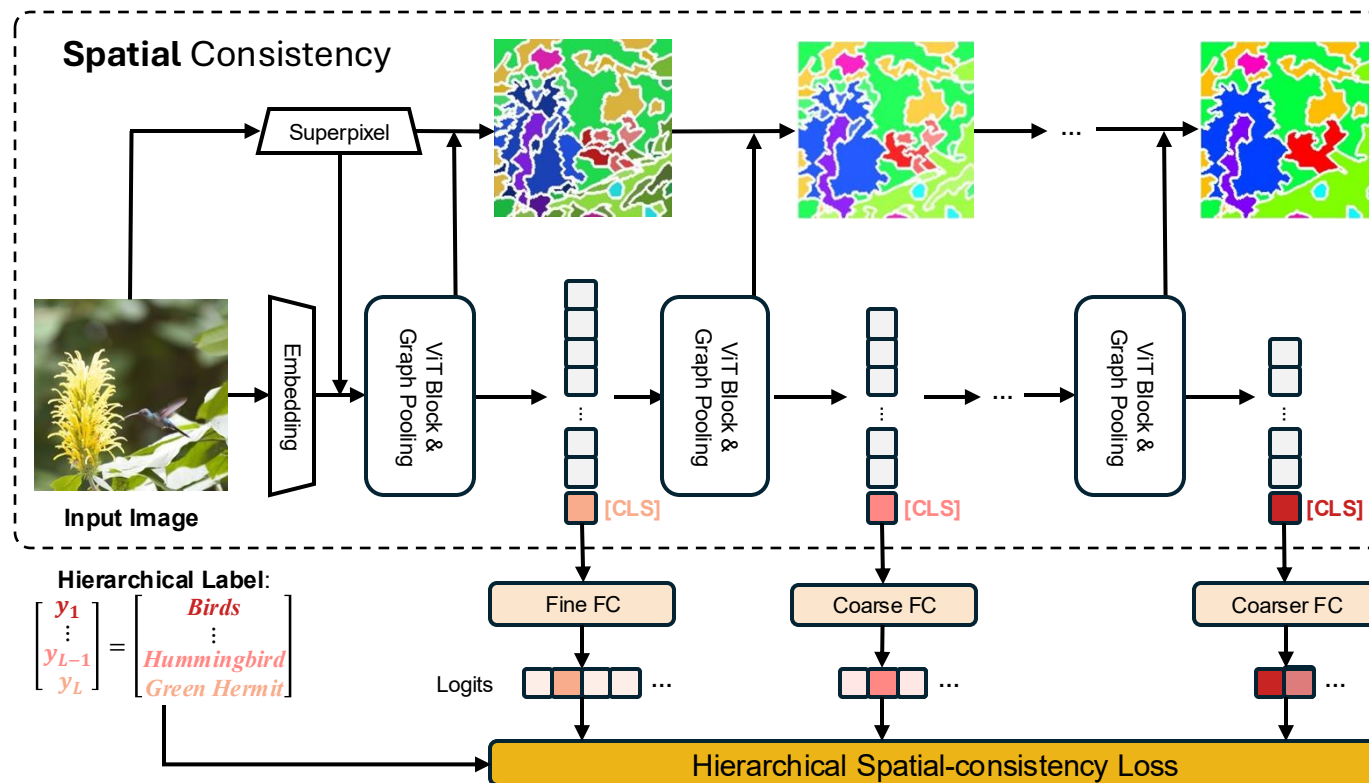
Visually Grounded Hierarchical Semantic Classification



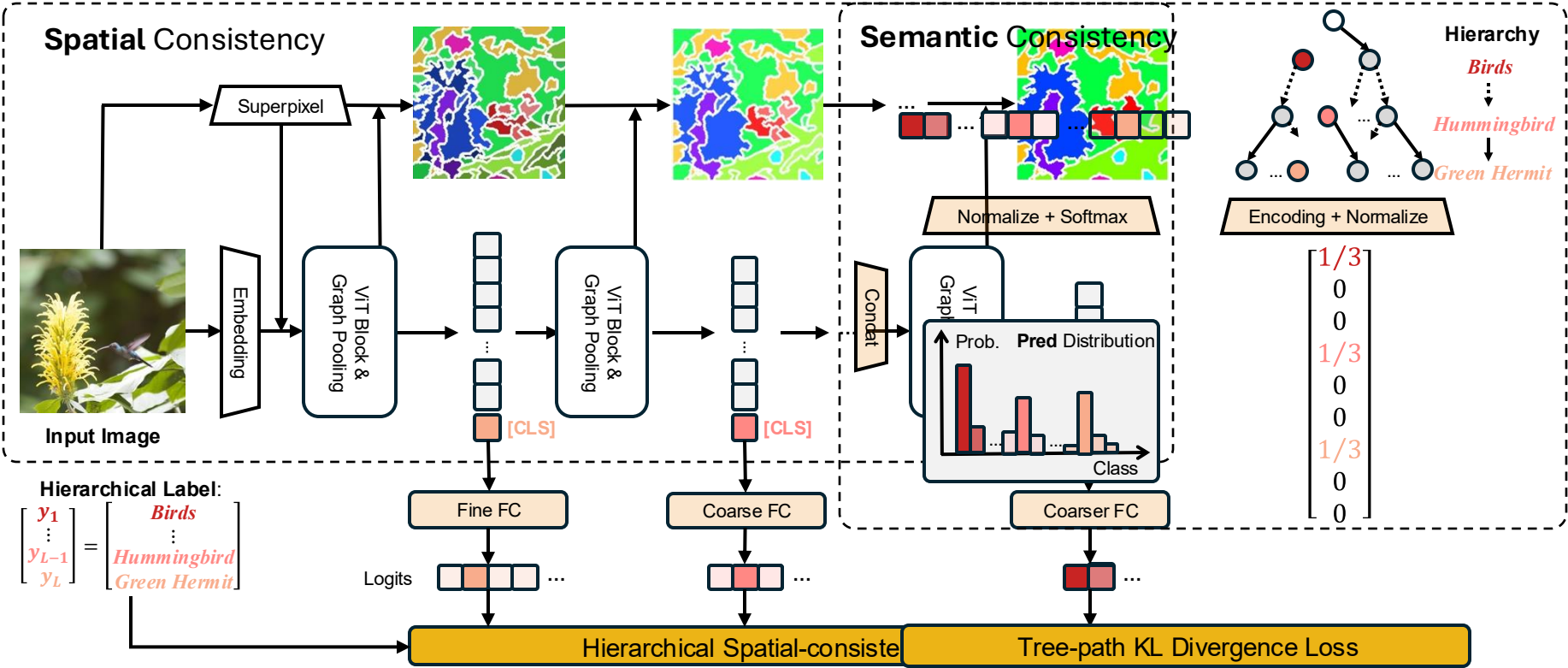
Overview



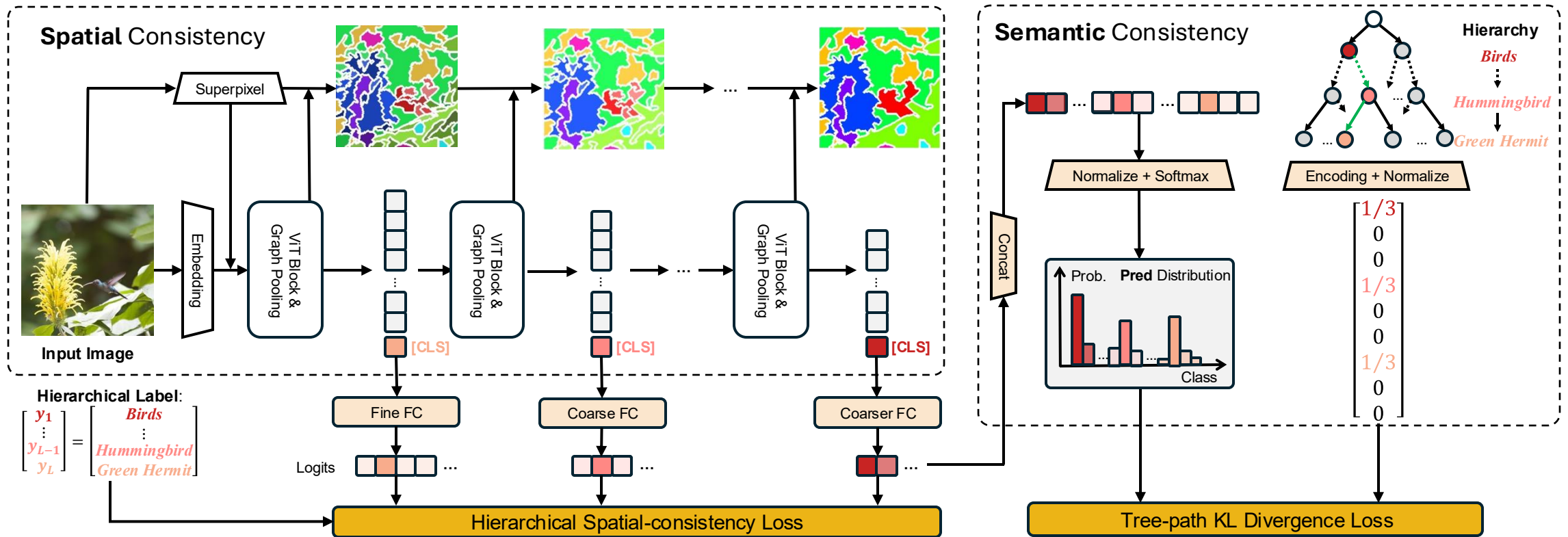
Novelty #1: H-CAST for Visual Consistency



Novelty #2: Tree-path KL Divergency Loss for Semantic Consistency

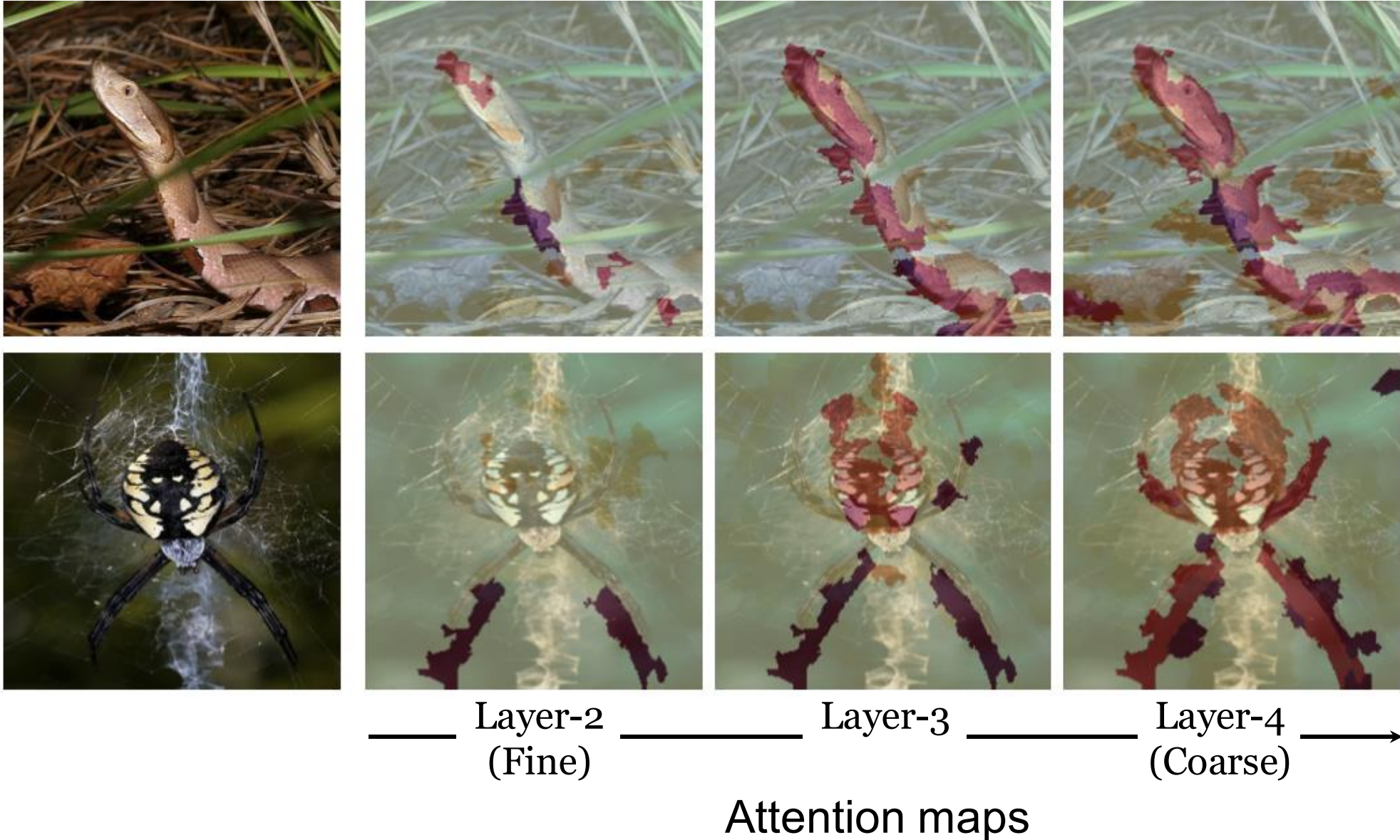


Novelty #2: Tree-path KL Divergency Loss for Semantic Consistency



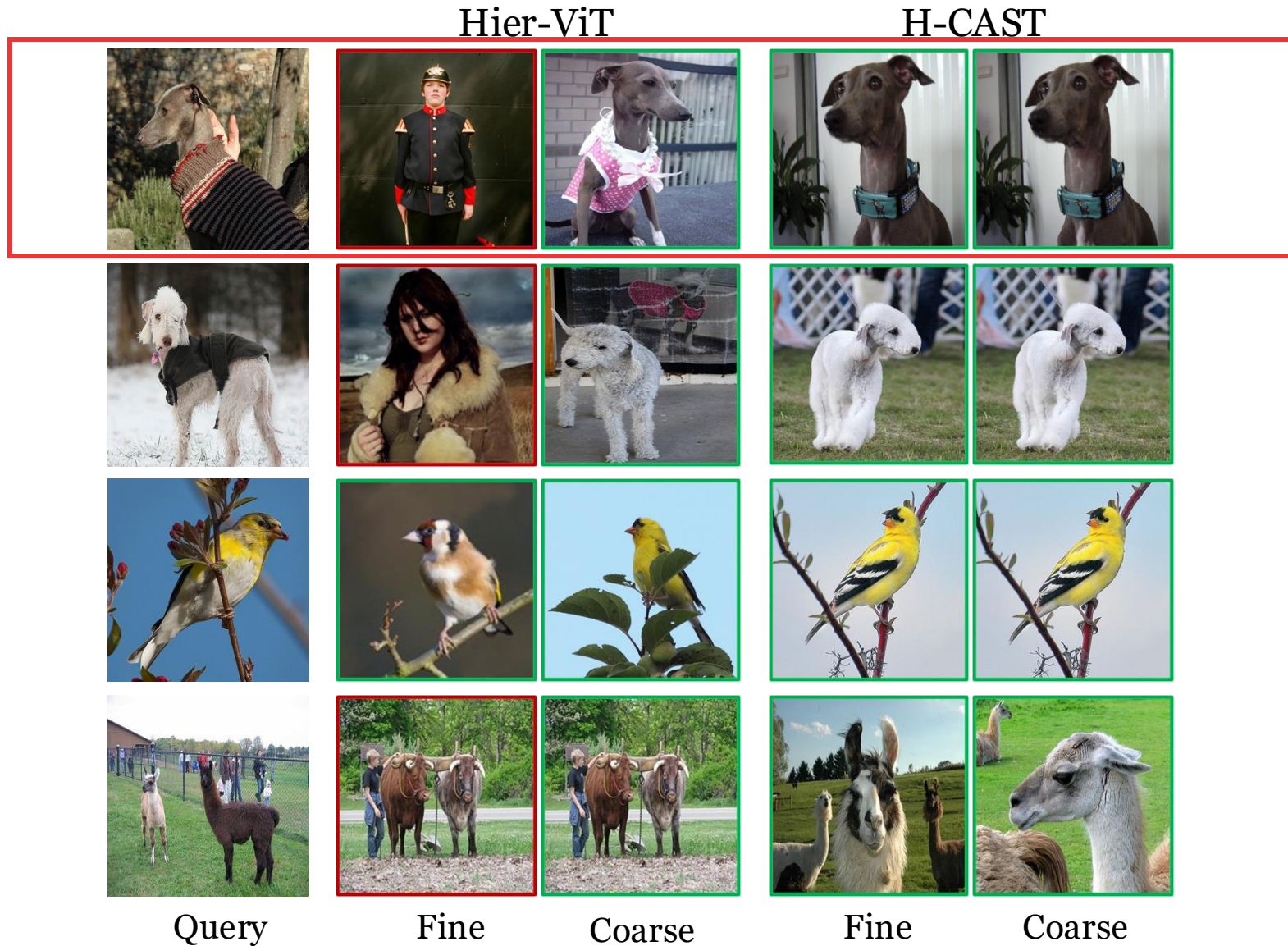
Results

1. H-CAST learns visual consistency from fine to coarse levels.



Results

1. H-CAST learns visual consistency from fine to coarse levels.
2. H-CAST delivers consistent representations.

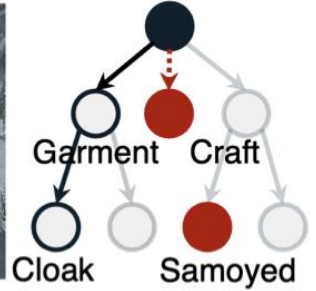
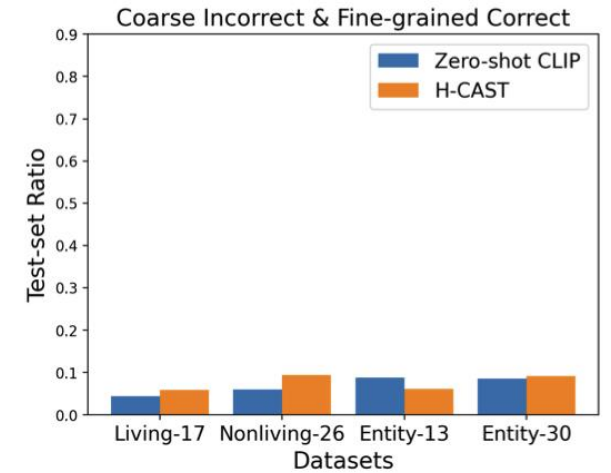
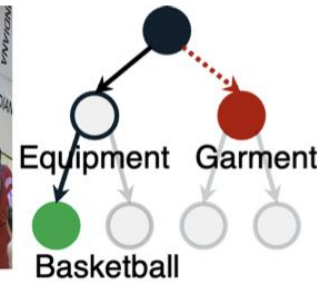
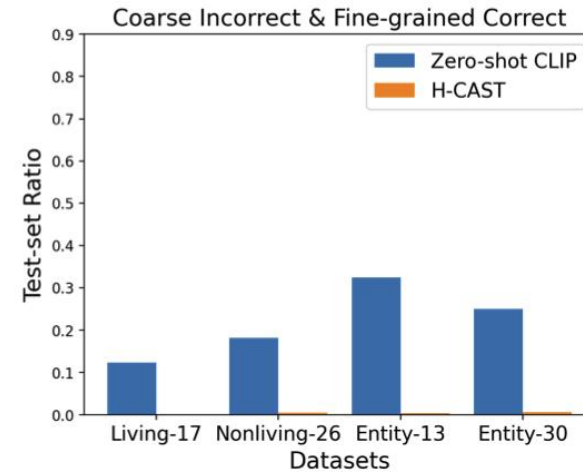
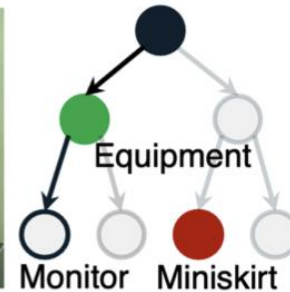
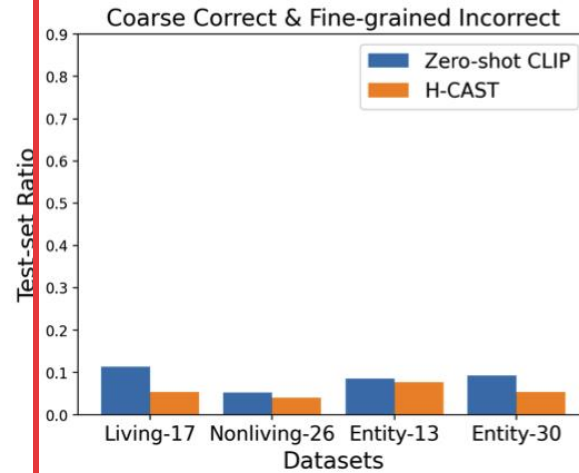
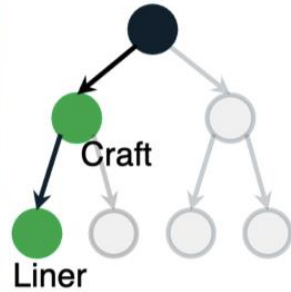
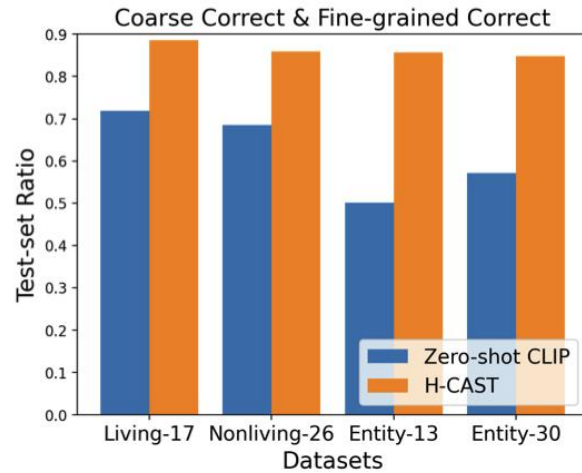


Results

1. H-CAST learns visual consistency from fine to coarse levels.
2. H-CAST delivers consistent representations.
3. H-CAST outperforms SOTA by 6%, ViT by 11%.

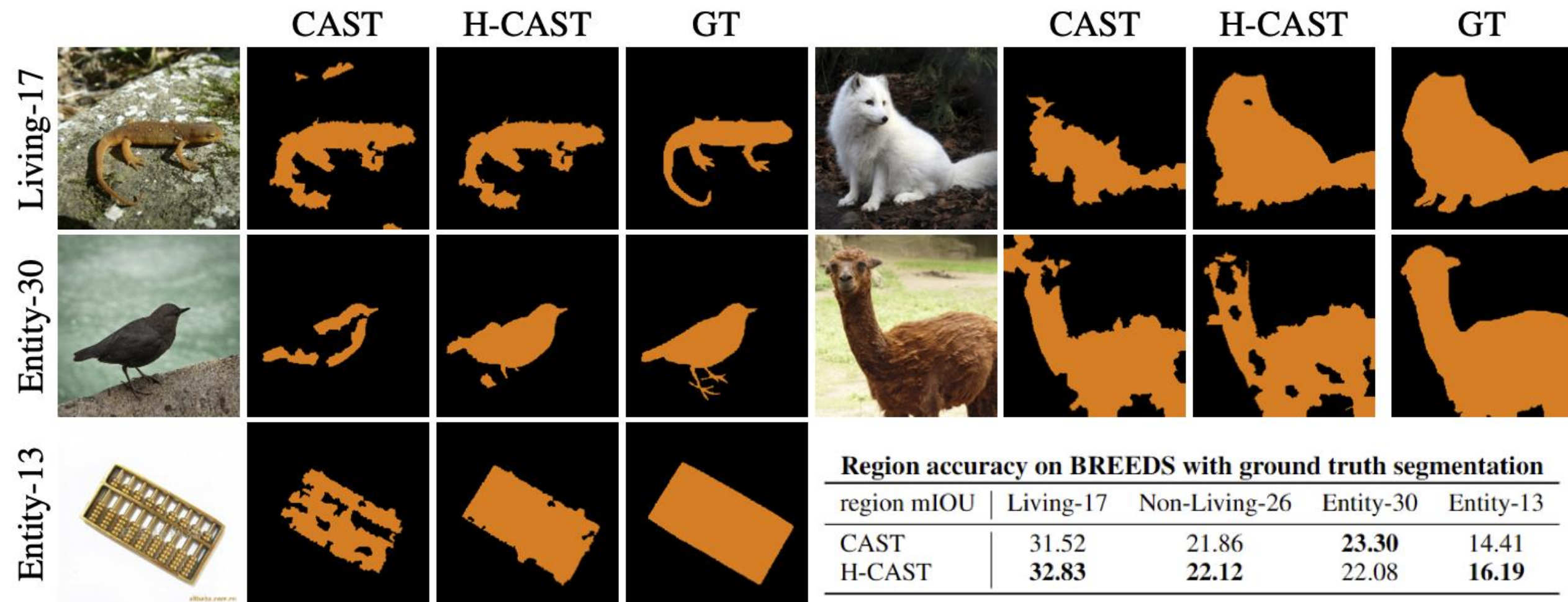
		Backbone	# Params	Living-17 (17-34)					Non-Living-26 (26-52)				
				FPA	Coarse	Fine	wAP	TICE	FPA	Coarse	Fine	wAP	TICE
Flat	Flat-ViT	ViT-S	65.0M	66.24	75.71	72.06	73.28	17.11	57.46	67.50	65.73	57.46	23.27
	Flat-ViT + HiE	ViT-S	65.0M	67.59	75.71	71.35	72.81	9.88	59.73	67.50	65.31	66.04	13.69
	Flat-CAST	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	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	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	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	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
	<i>Our Gains over SOTA</i>			+5.94	+3.29	+3.77	+3.61	+3.10	+6.36	+5.51	+3.00	+3.83	+4.28
		Backbone	# Params	Entity-13 (13-130)					Entity-30 (30-120)				
				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	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	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	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	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	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
	<i>Our Gains over SOTA</i>			+4.25	+3.42	+1.67	+2.62	+4.65	+4.98	+3.66	+2.10	+1.89	+5.81

- # Results:
1. H-CAST learns visual consistency from fine to coarse levels.
 2. H-CAST delivers consistent representations.
 3. H-CAST outperforms SOTA by 6%, ViT by 11%.
 4. H-CAST(Small) outperforms zero-shot CLIP(Base).



Results

1. H-CAST learns visual consistency from fine to coarse levels.
2. H-CAST delivers consistent representations.
3. H-CAST outperforms SOTA by 6%, ViT by 11%.
4. H-CAST(Small) outperforms zero-shot CLIP(Base).
5. H-CAST enhances weakly-supervised segmentation.



Thank you!

Code available at:

<https://github.com/pseulki/HCAST>

Visually Consistent Hierarchical Image Classification

Seulki Park

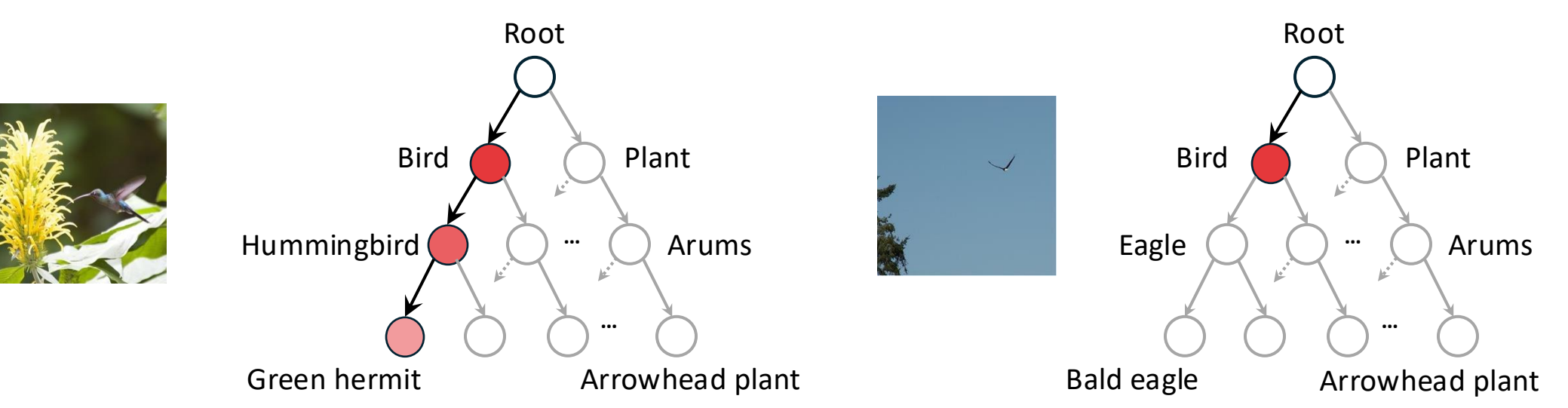
Youren Zhang

Stella X. Yu

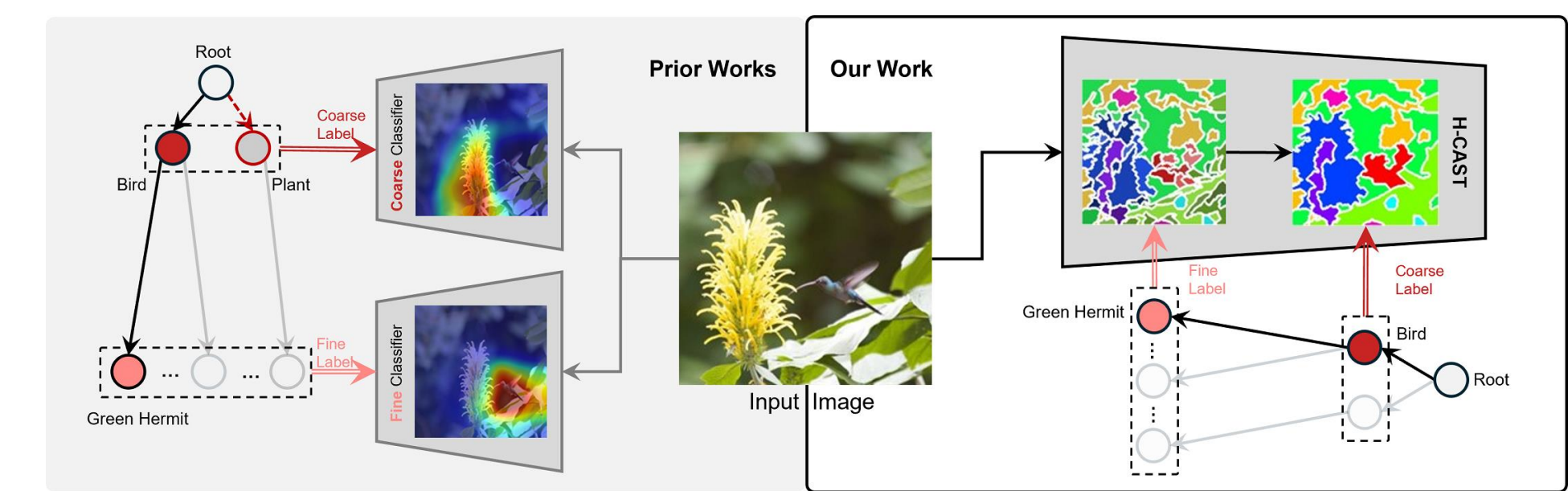
Sara Beery

Jonathan Huang

Hierarchical classification is necessary



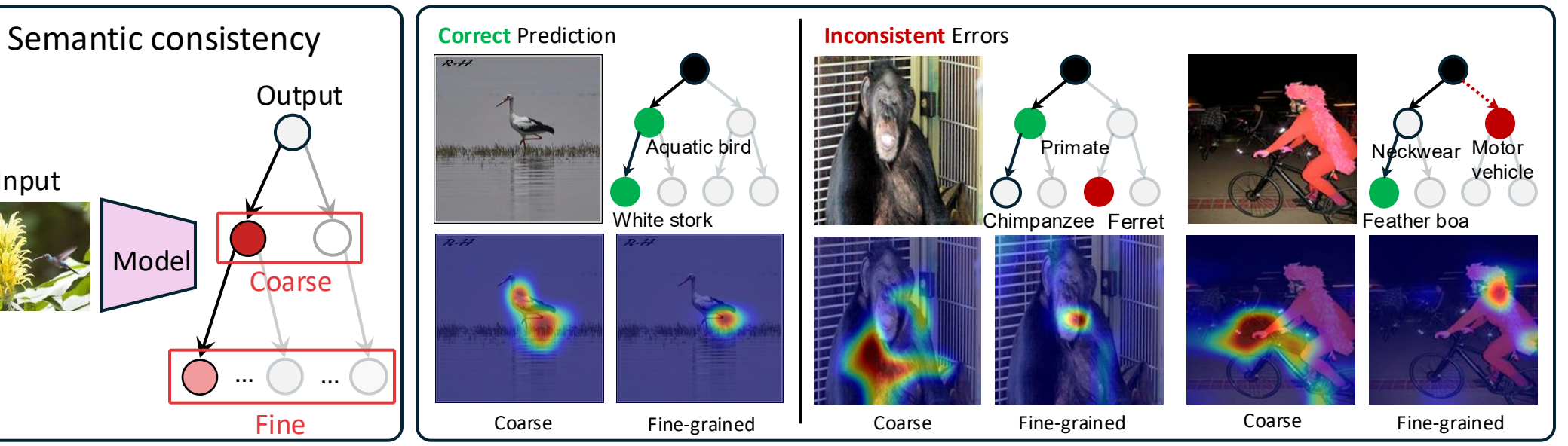
Internal visual consistency to align different levels



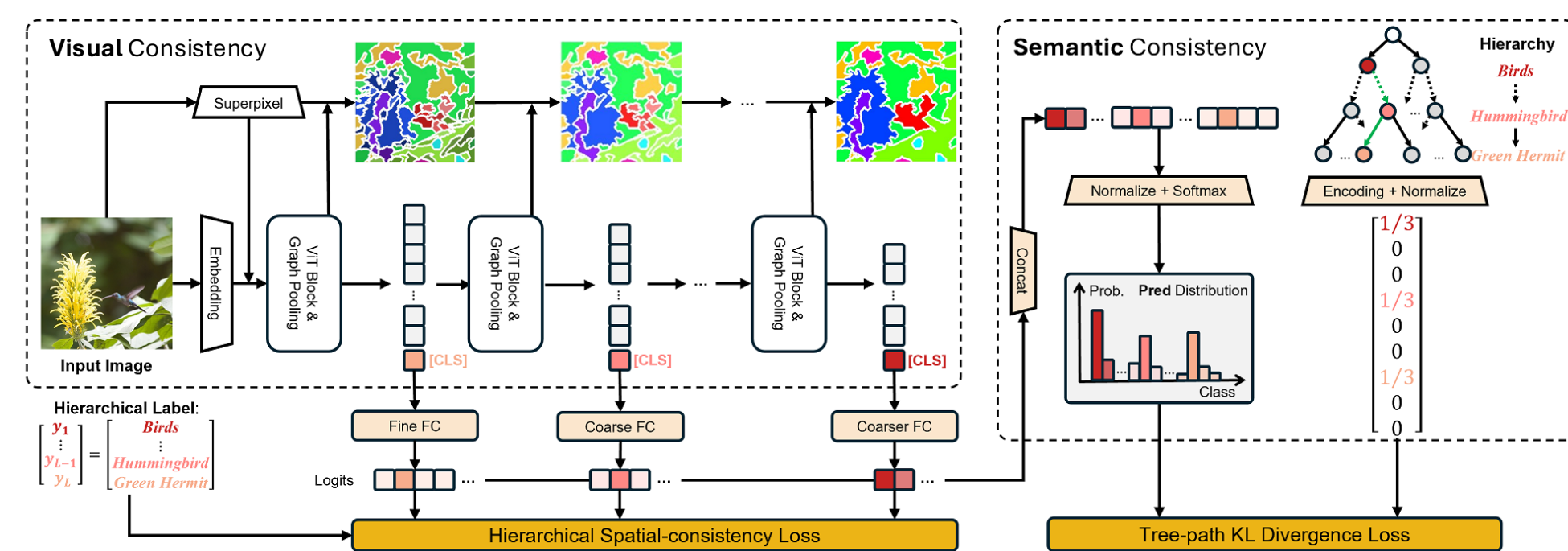
Result 2: H-CAST has consistent representations across



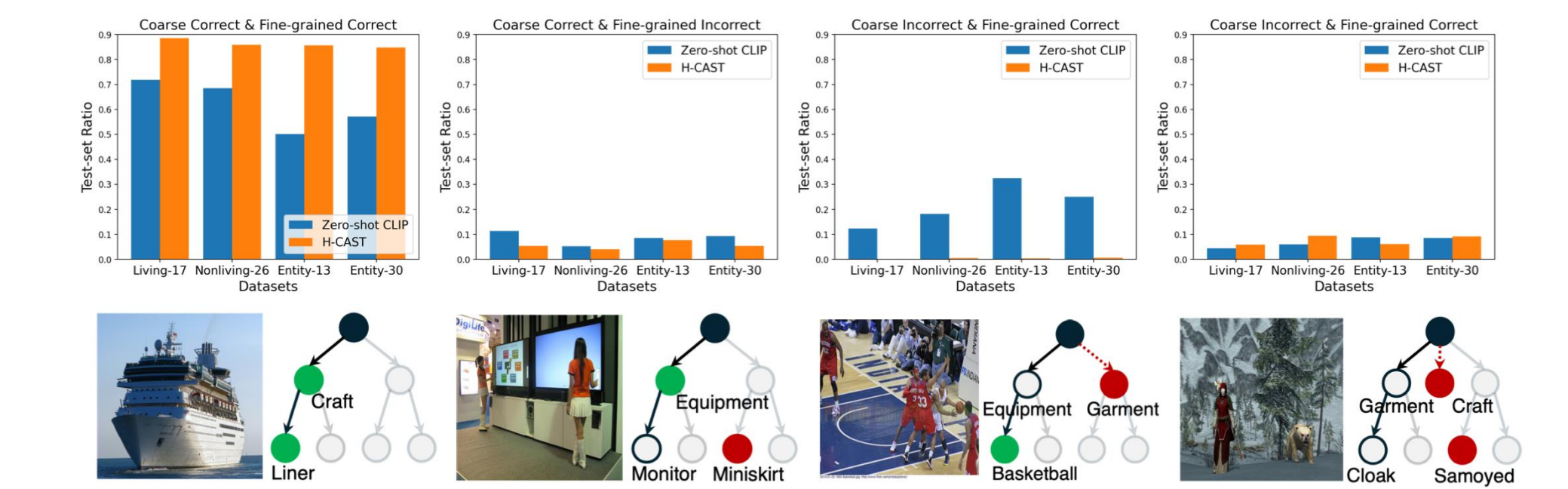
External semantic consistency is insufficient



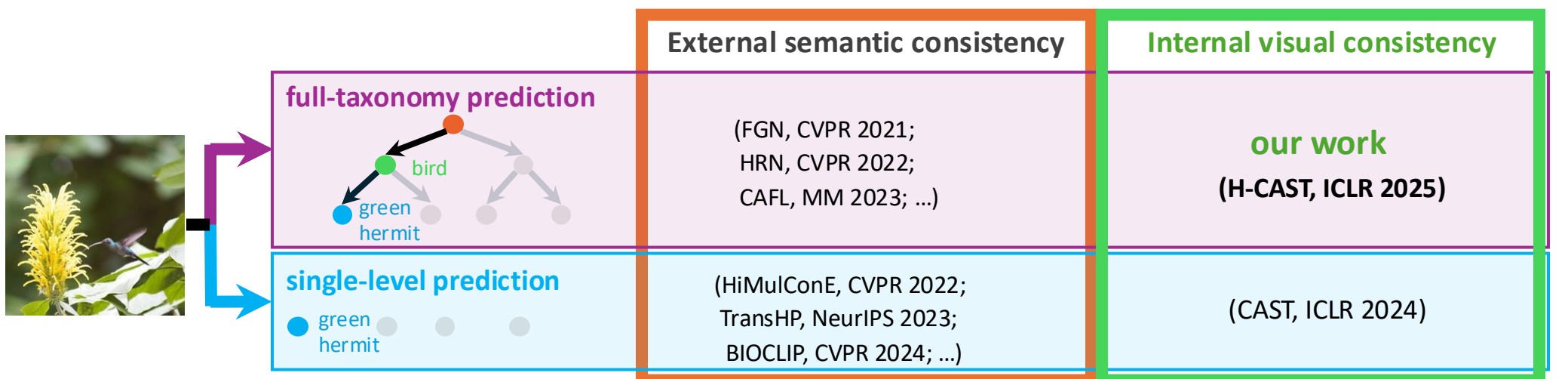
Our Model: visual consistency + semantic consistency



Result 3: H-CAST(small) beats zero-shot CLIP(base)



Contributions

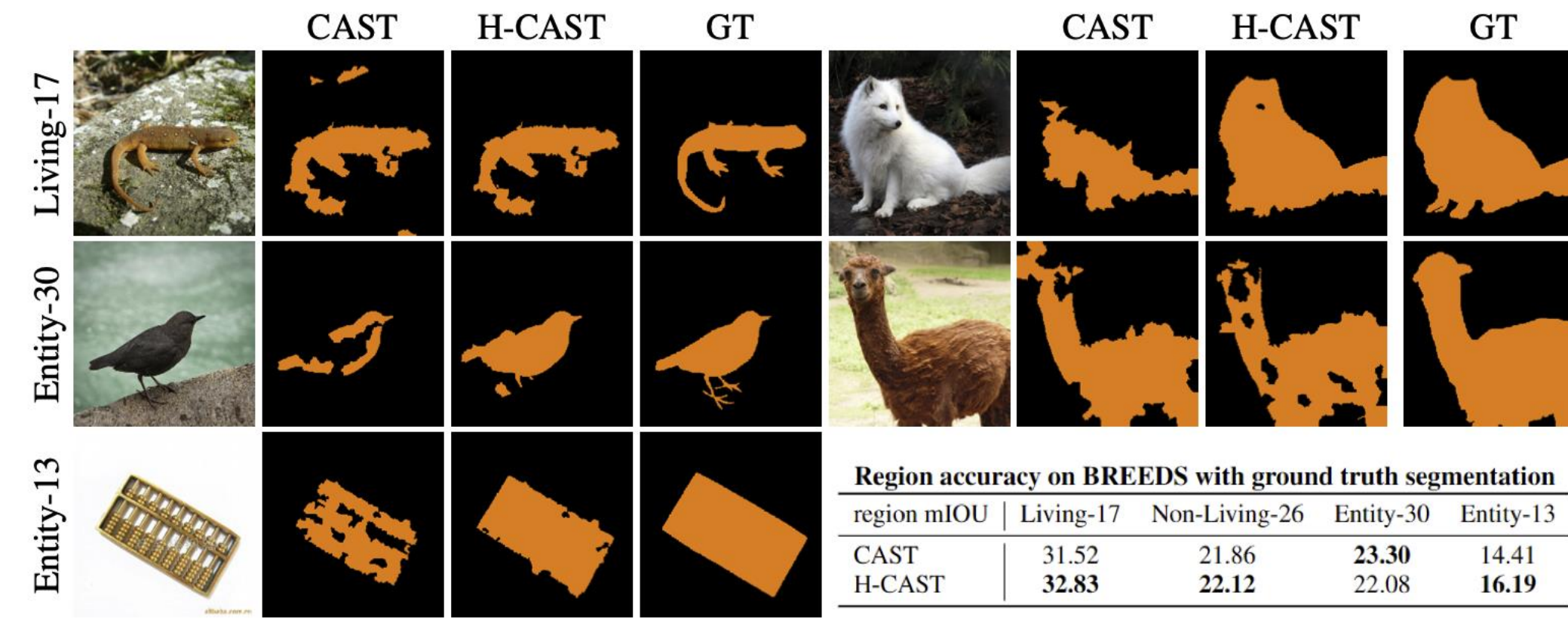


- First hierarchical classification model with internal visual consistency
- Better semantic consistency with tree-path KL divergence loss
- Internal visual consistency results in much higher accuracy and consistency
- Better unsupervised segmentation through hierarchical image classification

Result 1: H-CAST beats SOTA by 6% and ViT's by 11%

		Backbone	# Params	Living-17 (17-34)					Non-Living-26 (26-52)				
				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	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	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	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	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	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	1.69	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
		Backbone	# Params	Entity-13 (13-130)					Entity-30 (30-120)				
				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	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	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	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	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	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

Result 4: H-CAST enhances unsupervised segmentation



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