

Diversifying attention in HELM

A note on hierarchy-specific tokens for hierarchical multi-label classification of aerial images

Filip Josifljević¹ · Dubravko Ćulibrk^{2,3}

¹ Institute for Artificial Intelligence R&D of Serbia · ² Jožef Stefan Institute, Slovenia · ³ University of Novi Sad, Serbia

CONTEXT

One scene, many labels, few annotations

Multi-label

A marina tile is *dock*, *ship* and *water*, and also *Water Bodies* and *Inland Waters*.

Hierarchical

Labels form a CORINE-style tree: 4 top-level, 9 intermediate and 17 leaf categories.

Label-scarce

Annotating every level takes time and expertise, so semi-supervised learning is the practical route.

Flat multi-label classifiers treat all labels as independent and ignore this structure.

HELM models the label hierarchy explicitly

1 · CLASSIFICATION

Hierarchy tokens

A ViT-B gets $M = 30$ extra CLS tokens, one per node. Pooled, they feed a linear classifier.

2 · GRAPH LEARNING

GraphSAGE

The same tokens are node features on the label tree; the loss uses labelled samples only.

3 · SELF-SUPERVISION

BYOL

Patch tokens from two augmented views let the model learn from unlabelled images.

Claim in the original paper: the tokens “evolve into semantically meaningful representations that represent specific categories.”

THE QUESTION

Do the 30 tokens actually specialise?

A recent reproduction study (Josifljević, 2026) found:

- 1 Permanent symmetry.** Copying the pretrained CLS token 30 times with the same positional embedding gives 30 identical tokens at every layer. Fix: independent Gaussian noise, $\sigma = 0.02$.
- 2 A routing loophole.** Tokens sent most of their attention to each other rather than to the image.
- 3 A proposed remedy:** an attention-diversity loss.

This note: does the remedy hold across datasets, and does it give spatial specialisation?

Diversity loss and a data-driven hierarchy

Attention-diversity loss

$$L_{\text{div}} = \text{mean } \cos(a_i, a_j)$$

- a_i : token i 's attention over the image patches
- averaged over all token pairs $i \neq j$
- added to the HELM loss as $+\lambda \cdot L_{\text{div}}$

Hierarchy from the label matrix

- Y is an ancestor of X if X 's positive images are a strict subset of Y 's
- Parent = the smallest such superset
- Recovers the published UCM-HMLC tree: $4 + 9 + 17 = 30$ nodes
- Edge cases: co-occurring siblings, identical sets; resolved with CORINE

EXPERIMENTAL SETUP

Three benchmarks, 10% labels, three seeds

Dataset	Images (train / test)	Label hierarchy
UCM-HMLC	2,100 (1,667 / 433)	30 nodes: 4 / 9 / 17
AID	3,000 (2,400 / 600)	Same 30 nodes as UCM
DFC15	3,342 (2,673 / 669)	8 nodes, 2 levels (ours)*

ViT-B, supervised ImageNet weights

100 epochs · AdamW · cosine, lr 1e-4

Batch 16 · one V100

Macro AUPRC + ranking loss

* The original 17-node DFC15 hierarchy was not released. Ours is reproducible but not directly comparable.

Better on UCM-HMLC, not on AID or DFC15

Dataset	Baseline	+ Div, $\lambda = 0.1$	+ Div, $\lambda = 0.7$
UCM-HMLC	0.936 ± 0.013	0.932 ± 0.014	0.940 ± 0.008
AID	0.872 ± 0.016	0.863 ± 0.012	0.846 ± 0.011
DFC15	0.911 ± 0.009	0.899 ± 0.007	0.893 ± 0.011

UCM-HMLC

$\lambda = 0.7$ is best; but not significant

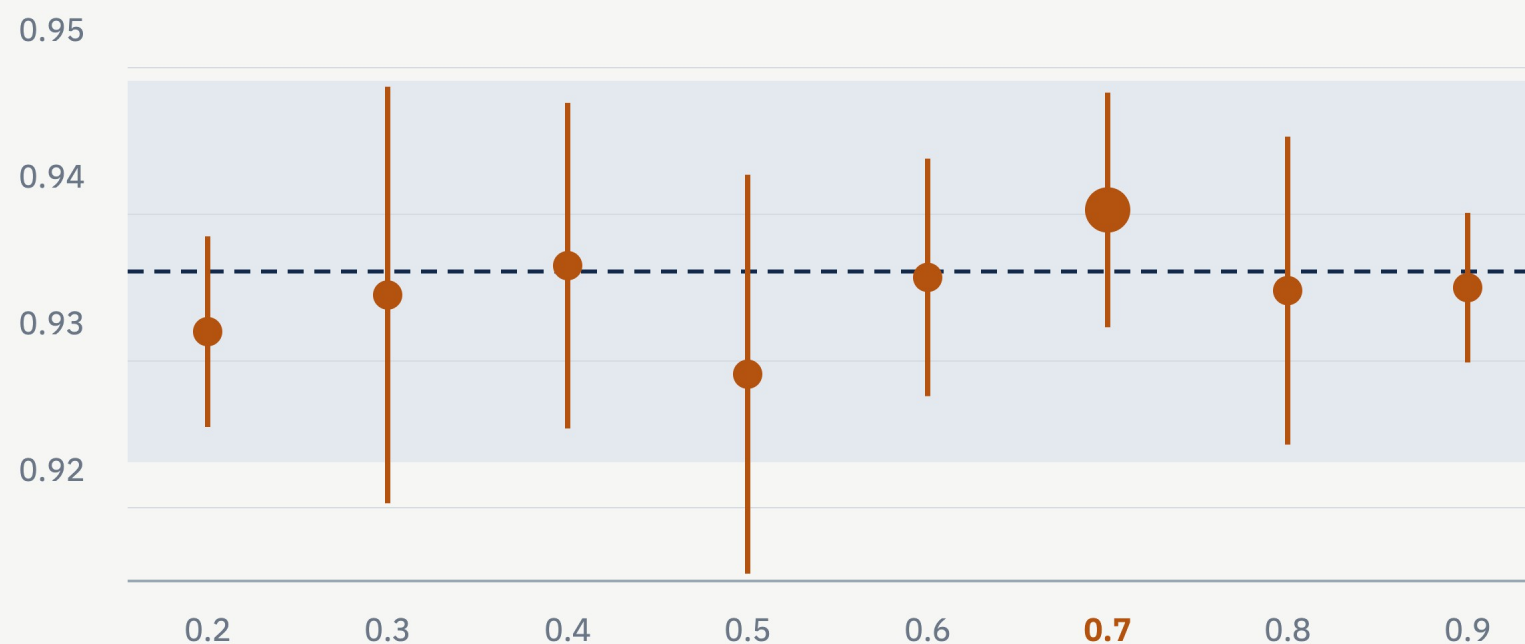
AID, DFC15

Baseline ahead; not significant

Ranking loss

Same pattern on all three

Only $\lambda = 0.7$ beats the baseline mean



$\lambda = 0.7$: 0.940 ± 0.008
vs baseline 0.936 ± 0.013

Every λ falls inside the baseline's ± 1 sd band.

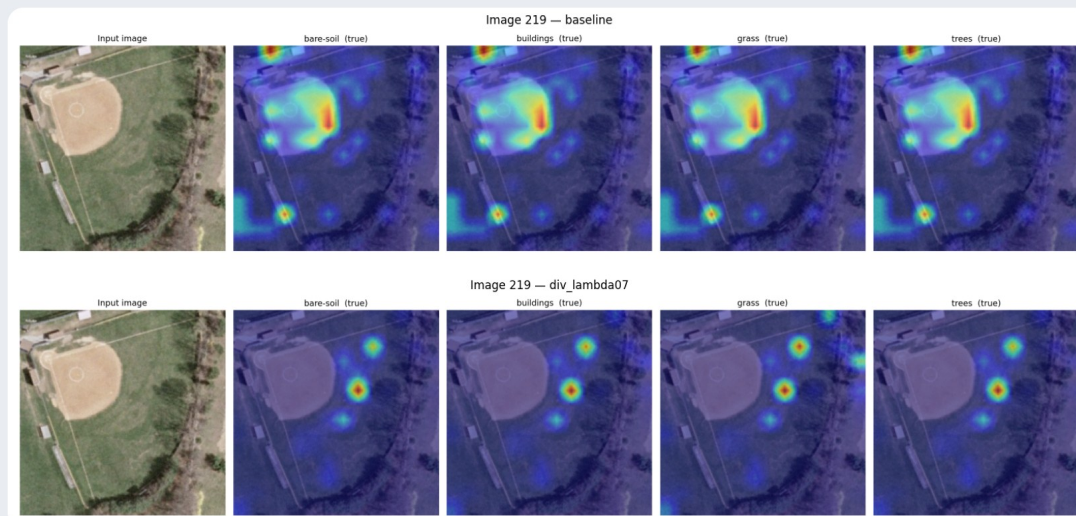
Dots, bars: diversity loss, mean $\pm$ sd
Dashed line, band: baseline, mean $\pm$ sd

Tokens now look at the image, at one spot

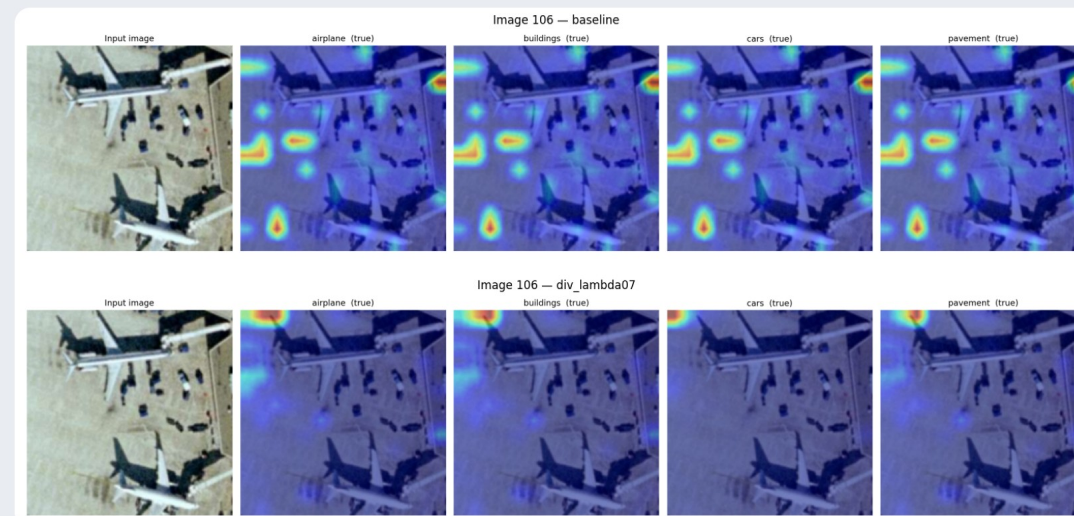


The routing loophole is gone: tokens put most attention on image patches. But *cars* and *trees* share one hotspot; pairwise cosine similarity is at its maximum.

The hotspot moves, but never splits

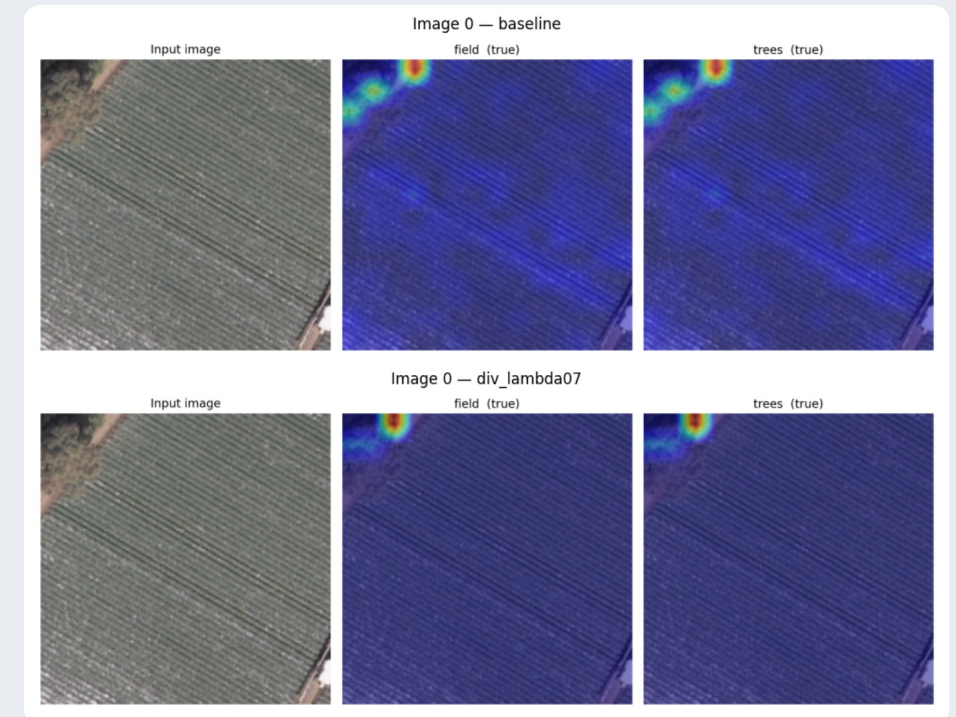
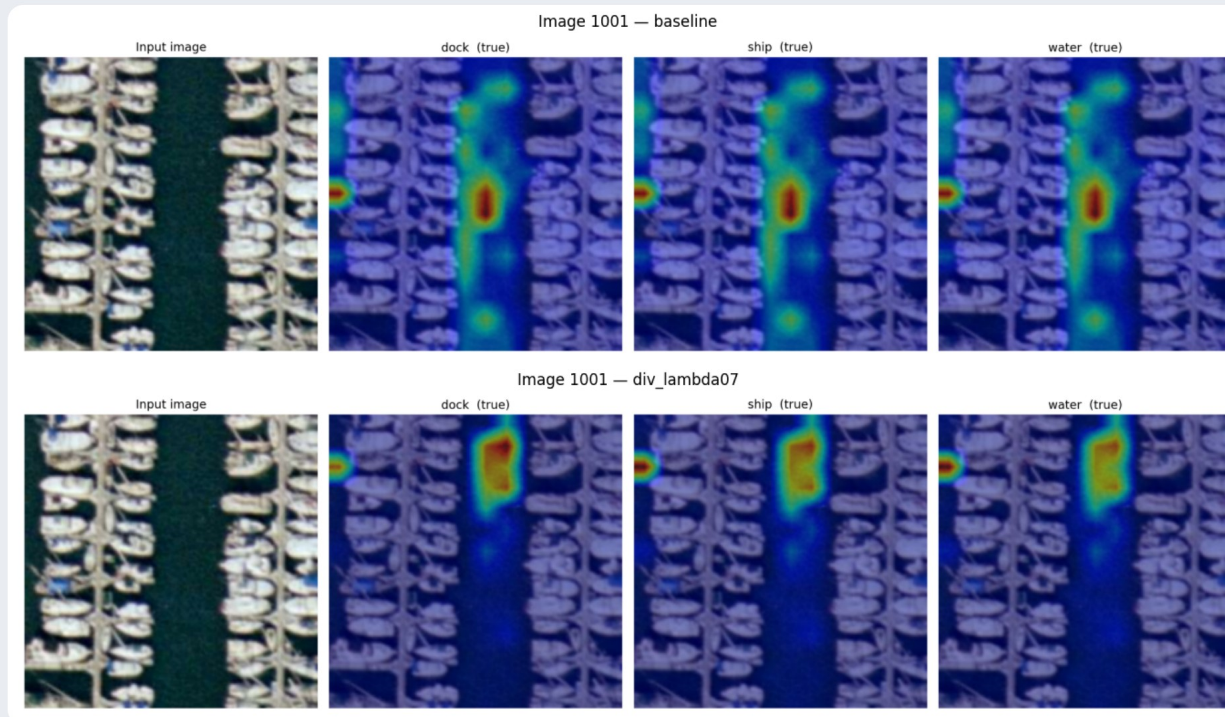


Top row: baseline. **Bottom row:** $\lambda = 0.7$. One map per true label.



$\lambda = 0.7$ changes *where* the tokens look, but within a row all maps match.

Sharper maps, still one map per image



$\lambda = 0.7$ gives more compact maps (marina: the upper channel; field: the tree corner only). Every label still gets the same map.

Example predictions: $\lambda = 0.7$ vs baseline



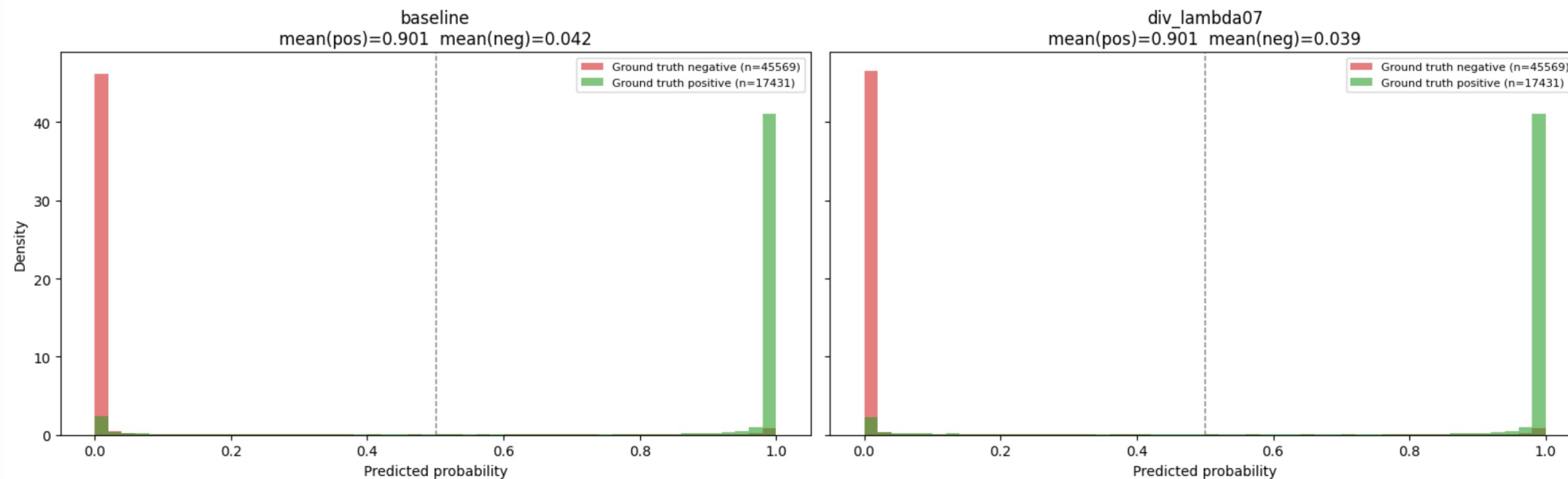
Image	Baseline	$\lambda = 0.7$
895 road	3 ✓ / 3 ✕	10 ✓ / 0 ✕
1040 marina	2 ✓ / 3 ✕	7 ✓ / 0 ✕
36 field	0 ✓ / 8 ✕	3 ✓ / 5 ✕
1603 river	8 ✓ / 0 ✕	14 ✓ / 0 ✕

✓ correct labels predicted, ✕ wrong labels predicted. Ground truth: 10, 7, 3 and 15 labels.

Exact label sets on two of four; the field is still confused with transport scenes.

Four examples only; the aggregate gain is small (slide 7).

Probability outputs barely change



Mean p on positives

0.901 → 0.901

Mean p on negatives

0.042 → 0.039

Both models are confidently bimodal; the loss trims false-positive mass slightly.

TAKEAWAYS

Distinct tokens, not distinct views

- 1 With noise initialisation, the tokens learn distinct representations and attend to the image.
- 2 The diversity loss is not a consistent win: better on UCM-HMLC, no significant gain on AID or DFC15.
- 3 Neither variant makes tokens attend to different regions. Spatial specialisation needs another mechanism.

Code and label hierarchies:

github.com/FilipJosifljjevic/hierarchical-remote-sensing

Thank you