HYPERRS: HYPERBOLIC EMBEDDINGS FOR HIERARCHICAL AERIAL IMAGE CLASSIFICATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Remote sensing scenes are naturally organized into semantic hierarchies, yet existing remote sensing foundation models typically discard this structure and learn flat representations in Euclidean space. Consequently, their predictions may be accurate without reflecting the semantic relations among categories. We introduce HyperRS, a hierarchical fine-tuning framework that explicitly aligns a pretrained hyperbolic image–text model with a remote sensing taxonomy. HyperRS applies hierarchical supervised contrastive learning at every taxonomy level and introduces a ranking loss that orders category embeddings by semantic granularity, placing fine-grained categories farther from the hyperbolic origin than their ancestors. Together with an entailment objective, these losses organize both the semantic and radial structure of the representation. Experiments on Million-AID and RSI-CB256 show that HyperRS outperforms the evaluated single-level and hierarchical models, achieving hierarchical precision of 87.2% and 94.9%, respectively. Detailed analysis shows that HyperRS preserves correct coarse-level predictions more often when fine-grained classification fails and recovers ancestor chains more faithfully through geometric traversal than the compared baselines. It also maintains higher hierarchical precision than baselines across three common remote sensing image corruptions.

1 INTRODUCTION

Remote sensing scenes are naturally understood at multiple levels of semantic granularity. An image labeled as an *apron*, for example, belongs to the broader category *airport*, which is itself a type of *transportation land*. Such relations are explicitly encoded in widely used remote sensing taxonomies. Million-AID (Long et al., 2021), for instance, organizes aerial scenes into a multi-level hierarchy containing both coarse land-use concepts and fine-grained scene categories. This hierarchy provides more than additional labels: it specifies which categories are semantically related and how their meanings change with granularity. Nevertheless, most representation-learning methods use only flat category supervision and therefore discard much of this structure.

The rapid development of vision–language foundation models has further advanced remote sensing scene understanding. Models such as RemoteCLIP and GeoRSCLIP (Liu et al., 2024; Zhang et al., 2024), following CLIP (Radford et al., 2021; Jia et al., 2021), embed images and text into a shared Euclidean space and learn representations through contrastive alignment. These models are effective at measuring semantic similarity, but their training objectives do not explicitly encode ancestor–descendant relations or distinguish different levels of abstraction. In particular, normalized Euclidean embeddings place all concepts on the same unit hypersphere, providing no direct geometric representation of semantic depth. A model may therefore classify individual levels accurately while organizing related categories inconsistently across levels.

Hyperbolic geometry offers an inductive bias suited to this structure: its exponential volume growth accommodates tree-like branching, while radial distance can represent increasing semantic specificity (Nickel & Kiela, 2017; Lee, 2018). MERU (Desai et al., 2023) demonstrates that such organization can emerge during hyperbolic image–text pretraining, but its training does not explicitly align representations with a given taxonomy. This leaves a gap between the hierarchical capacity of the embedding space and the taxonomic supervision available in remote sensing. We address this gap by jointly learning category semantics across levels and explicitly organizing their geometric order.

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

We introduce HyperRS, a hierarchical fine-tuning framework built on pretrained MERU (Desai et al., 2023). A *hierarchical supervised contrastive loss* (HSC) applies supervised contrastive learning (Khosla et al., 2020) at every annotated taxonomy level, aligning images and label texts under both coarse- and fine-grained semantics, allowing the representation to preserve information across the complete taxonomy. While this objective learns semantic associations across levels, it does not explicitly prescribe their radial order. We therefore introduce a *ranking loss* that requires each fine-grained category to lie farther from the origin than its parent by a predefined margin. This constraint explicitly aligns radial distance with semantic specificity. The resulting representation encodes complementary information along two dimensions: relative proximity captures semantic similarity, while radial position captures semantic granularity.

Experiments on Million-AID and RSI-CB256 show that HyperRS improves hierarchy classification performance over the evaluated single-level and hierarchical models, achieving hierarchical precision of 87.2% and 94.9%, respectively. On both datasets, it preserves correct coarse-level predictions more often when fine-grained classification fails. Its geometric traversals cover the complete ground-truth ancestor chain at a leaf-macro-averaged rate of 36.4%, compared with 22.6% for MERU trained with the same hierarchical supervision and at most 2.9% for the evaluated Euclidean baselines. The advantage in hierarchical precision also persists under the tested image corruptions.

Summary of Contributions

- We propose HyperRS, a hierarchical fine-tuning framework that injects explicit multi-level taxonomic supervision into a pretrained hyperbolic image–text foundation model. To the best of our knowledge, this is the first framework to explicitly align a large-scale hyperbolic vision–language representation with a remote sensing taxonomy.
- We develop a hierarchical supervised contrastive objective that learns jointly from every annotated taxonomy level. Compared with the evaluated single-level and hierarchical models, it produces substantially stronger predictions across the hierarchy.
- We introduce a radial ranking loss that encodes semantic granularity. Combined with hierarchical contrastive and entailment objectives, the resulting framework achieves more coherent hierarchical predictions, stronger ancestor-chain recovery, and improved robustness under common image corruptions.

2 Related Work

2.1 Remote Sensing Foundation Models

Motivated by the success of contrastive vision-language pretraining in natural images (Radford et al., 2021; Jia et al., 2021), a growing body of work has developed foundation models for remote sensing. Contrastive vision-language models, including RemoteCLIP (Liu et al., 2024) and GeoRSCLIP (Zhang et al., 2024), align aerial imagery with textual descriptions to support open-vocabulary and zero-shot recognition, while masked-image-modeling approaches such as SatMAE (Cong et al., 2022) and Scale-MAE (Reed et al., 2023) learn transferable representations from unlabeled imagery for downstream adaptation. Both paradigms learn flat Euclidean representations. The image–text models compare L_2-normalized embeddings by cosine similarity, which places every concept on the same unit hypersphere and leaves no geometric dimension for semantic depth; their contrastive objectives likewise encode no ancestor–descendant relations. As a result, the multi-level taxonomies that accompany remote sensing datasets such as Million-AID (Long et al., 2021) are not used when these foundation models are pretrained or adapted.

2.2 Hyperbolic Representation Learning

Hyperbolic space is well suited to hierarchical data: its volume grows exponentially with distance from the origin, making it a continuous analogue of trees (Nickel & Kiela, 2017; Lee, 2018). It has been used to embed taxonomies (Nickel & Kiela, 2017; 2018) and, more recently, to learn visual representations (Khrulkov et al., 2020; Atigh et al., 2022; Ermolov et al., 2022). For vision–language models, MERU (Desai et al., 2023) and HyCoCLIP (Pal et al., 2025) adapt the hyperbolic entailment cones of Ganea et al. (2018) to large-scale image–text training, placing generic concepts near the origin and specific ones farther away. Their hierarchy, however, emerges from image–text pairs

rather than from a given taxonomy. HypStructure (Sinha et al., 2024) does use a label hierarchy, but as a regularizer for image classifiers. HyperRS instead aligns a pretrained hyperbolic image–text model with a multi-level taxonomy and explicitly orders its label embeddings by radial distance.

2.3 Hierarchical Classification and Evaluation

Remote sensing taxonomies have been exploited by cascaded CNN classifiers (Sen & Keles, 2022), multi-task networks that jointly predict coarse and fine labels (Zeng et al., 2022), and hierarchical prototypical networks (Hamzaoui et al., 2022). Outside remote sensing, Zhang et al. (2022) add hierarchy-preserving penalties to a supervised contrastive loss (Khosla et al., 2020). In these methods the hierarchy constrains predictions or pairwise similarities, but the Euclidean representations have no geometric notion of depth, so a concept's granularity is not reflected in where it is embedded.

Hierarchical predictions are commonly evaluated with hierarchical precision (Kiritchenko et al., 2005; 2006; Kosmopoulos et al., 2015), the Wu–Palmer similarity (Wu & Palmer, 1994), and the exact-match ratio over all levels (Zhang & Zhou, 2013). These metrics are averaged over all samples and do not single out the samples a model misclassifies at the finest level, where it matters whether the coarse prediction survives. They also evaluate only predictions and not the embedding geometry, although MERU (Desai et al., 2023) shows qualitatively that an image's ancestors can be recovered by traversing its embedding toward the origin. We therefore measure whether errors stay within the correct subtree and quantify how well traversals recover the taxonomy.

3 Method

HyperRS fine-tunes a pretrained hyperbolic image-text backbone under explicit hierarchical supervision. Given an image paired with its class label, the image and text encoders each produce a Euclidean feature, which is then lifted onto the Lorentz hyperbolic space via the exponential map at the origin. We initialize both encoders from the pretrained MERU (Desai et al., 2023) checkpoint and fine-tune with three objectives: a hierarchical supervised contrastive loss (HSC, Sec. 3.2), an entailment loss (Ganea et al., 2018; Desai et al., 2023) (Sec. 3.1) and a ranking loss (Sec. 3.3). An overview of the overall pipeline is shown in Fig. 1. We first briefly recall the Lorentz-model preliminaries, then introduce our two hierarchical objectives.

3.1 Preliminaries: The Lorentz Model

We build on MERU (Desai et al., 2023) and adopt the Lorentz model of n-dimensional hyperbolic space with a learnable curvature $-c$ ($c > 0$). A point $\mathbf{x} = [x_{\text{time}}, \mathbf{x}_{\text{space}}] \in \mathbb{R}^{n+1}$ has a time component $x_{\text{time}} \in \mathbb{R}$ and a space component $\mathbf{x}_{\text{space}} \in \mathbb{R}^n$, and the Lorentzian inner product is

$$\langle \mathbf{x}, \mathbf{y} \rangle_{\mathcal{L}} = \mathbf{x}_{\text{space}}^{\top} \mathbf{y}_{\text{space}} - x_{\text{time}}\, y_{\text{time}}. \tag{1}$$

The hyperbolic space is $\mathcal{L}^n = \{\, \mathbf{x} \in \mathbb{R}^{n+1} : \langle \mathbf{x}, \mathbf{x} \rangle_{\mathcal{L}} = -1/c,\ x_{\text{time}} > 0 \,\}$, and we take its origin $\mathbf{O} = [\, 1/\sqrt{c},\ \mathbf{0} \,]$ to represent the most generic concept. The Lorentzian distance between two points $\mathbf{x}, \mathbf{y} \in \mathcal{L}^n$ is

$$d_{\mathcal{L}}(\mathbf{x}, \mathbf{y}) = \sqrt{1/c}\, \cosh^{-1}\big(-c\, \langle \mathbf{x}, \mathbf{y} \rangle_{\mathcal{L}} \big). \tag{2}$$

Lifting Euclidean features onto the hyperbolic space. Given a Euclidean feature $\mathbf{v}_{\text{space}} \in \mathbb{R}^n$ produced by an encoder, we treat it as a tangent vector at the origin, $\mathbf{v} = [\, 0, \mathbf{v}_{\text{space}} \,] \in \mathcal{T}_{\mathbf{O}} \mathcal{L}^n \cong \mathbb{R}^n$, and map it onto the hyperbolic space with the exponential map $\exp_{\mathbf{O}} : \mathcal{T}_{\mathbf{O}} \mathcal{L}^n \to \mathcal{L}^n$. Since tangent vectors at $\mathbf{O}$ have zero time component, the Lorentzian norm reduces to the Euclidean norm $\|\mathbf{v}_{\text{space}}\|$, and following Desai et al. (2023) the map reduces to computing the space component

$$\mathbf{x}_{\text{space}} = \frac{\sinh\big(\sqrt{c}\, \|\mathbf{v}_{\text{space}}\|\big)}{\sqrt{c}\, \|\mathbf{v}_{\text{space}}\|}\, \mathbf{v}_{\text{space}}, \tag{3}$$

after which x_{time} is determined by the constraint $\langle \mathbf{x}, \mathbf{x} \rangle_{\mathcal{L}} = -1/c$, so that $\mathbf{x}$ always lies on $\mathcal{L}^n$. All image and text embeddings in HyperRS are obtained this way.

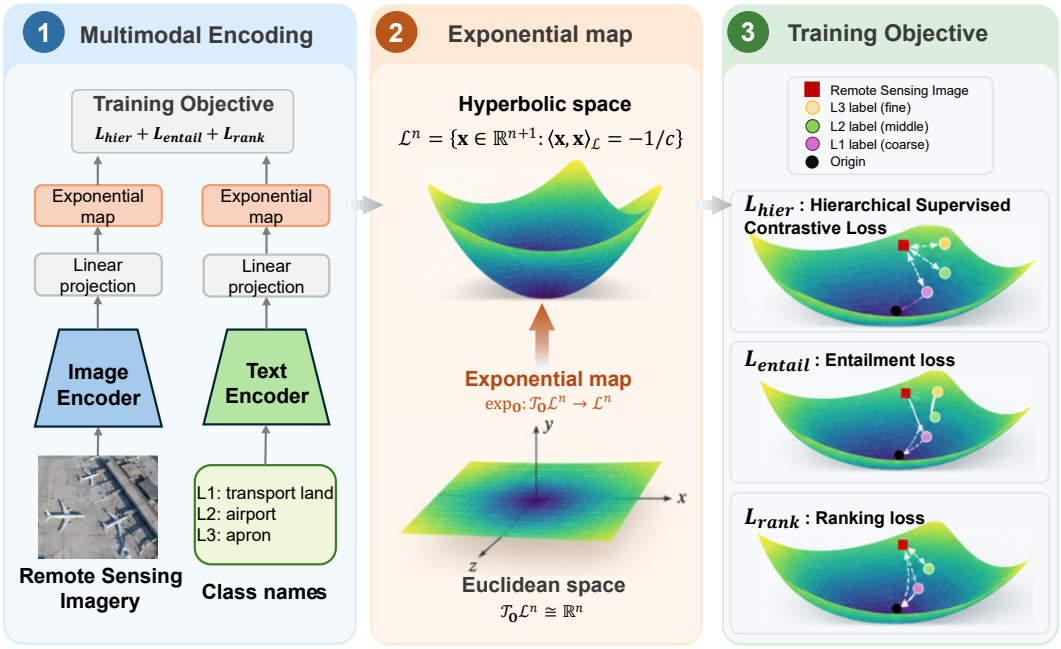

Figure 1: **Overview of HyperRS.** Image and text are encoded, lifted to the Lorentz hyperbolic space via the exponential map at the origin, and fine-tuned with three objectives: the hierarchical supervised contrastive loss L_{hier}, the ranking loss L_{rank}, and the entailment loss L_{entail}.

Entailment cone. To encode partial-order (*text-entails-image*) relations, each point $\mathbf{x}$ is associated with an entailment cone whose half-aperture narrows with distance from the origin,

$$\text{aper}(\mathbf{x}) = \sin^{-1}\left(\min\left\{\frac{2K}{\sqrt{c}\,\|\mathbf{x}_{\text{space}}\|},\ 1-\epsilon\right\}\right), \tag{4}$$

with a constant $K = 0.1$ and $\epsilon = 10^{-8}$.

The clamp keeps the aperture well defined near the origin: when $\sqrt{c}\,\|\mathbf{x}_{\text{space}}\| \le 2K$, the half-aperture is fixed at $\sin^{-1}(1-\epsilon) \approx \pi/2$.

For a pair $(\mathbf{x}, \mathbf{y})$, the exterior angle $\text{ext}(\mathbf{x}, \mathbf{y})$ measures how far $\mathbf{y}$ lies outside the cone projected from $\mathbf{x}$, and the entailment loss penalizes violations of the partial order,

$$L_{\text{entail}}(\mathbf{x}, \mathbf{y}) = \max\big(0,\ \text{ext}(\mathbf{x}, \mathbf{y}) - \text{aper}(\mathbf{x})\big). \tag{5}$$

We keep this per-pair loss with $\mathbf{x}$ a label text and $\mathbf{y}$ an image, and apply it between each image and its ancestor label at every annotated level (Sec. 3.4); the full expression for $\text{ext}(\cdot,\cdot)$ is given in Appendix A.

3.2 Hierarchical Supervised Contrastive Loss

Standard contrastive learning treats each image–text pair as the only positive. In a taxonomy, however, many images share a label at coarser levels, and counting their label texts as negatives introduces false repulsion. HSC therefore uses level-dependent positives. Consider a batch of B images with their class labels. Let $y_i^{(\ell)}$ be the level-ℓ label of image i and d_i its deepest annotated level; because the taxonomy has mixed depth, only the images in $\mathcal{V}_\ell = \{\, i : \ell \le d_i \,\}$ carry a level-ℓ label. The positives of image i at level ℓ are $P_i^{(\ell)} = \{\, j \in \mathcal{V}_\ell : y_j^{(\ell)} = y_i^{(\ell)} \,\}$, the in-batch images with the same level-ℓ label. We denote by $z_i \in \mathcal{L}^n$ the image embedding and by $t_i^{(\ell)} \in \mathcal{L}^n$ the embedding of its level-ℓ label text, formed by filling the label name into a prompt template (e.g., "an aerial image of {}") sampled from a fixed set. We define the image-to-text similarity $s_{ij}^{(\ell)} = -\tau\, d_{\mathcal{L}}(z_i, t_j^{(\ell)})$ as the

negative Lorentzian distance scaled by a learnable temperature τ, and the image-to-text supervised contrastive loss at level ℓ as

$$L_\ell^{i\to t} = \frac{1}{|\mathcal{V}_\ell|} \sum_{i\in\mathcal{V}_\ell} \frac{-1}{|P_i^{(\ell)}|} \sum_{j\in P_i^{(\ell)}} \log \frac{\exp\big(s_{ij}^{(\ell)}\big)}{\sum_{k\in\mathcal{V}_\ell} \exp\big(s_{ik}^{(\ell)}\big)}. \tag{6}$$

The text-to-image loss $L_\ell^{t\to i}$ is defined symmetrically by swapping the roles of images and texts.

3.3 Ranking Loss

The HSC objective aligns samples with their labels at each level, but it constrains only *relative* similarities and leaves the radial position of a concept unconstrained, i.e., its distance from the origin. We want distance from the origin $\mathbf{O}$ to encode specificity, with generic concepts near $\mathbf{O}$ and specific ones farther away, and therefore add a ranking loss that enforces this radial order. Let $r(z) = d_{\mathcal{L}}(z, \mathbf{O})$ be the Lorentzian distance of an embedding from the origin. Along each sample's ground-truth path, a child (finer) class should lie farther from the origin than its parent (coarser) class. Over each pair of consecutive levels ℓ and $\ell+1$ along sample i's annotated path of d_i levels, we penalize any case where the finer class is not at least a margin m farther from the origin:

$$L_{\text{rank}} = \frac{1}{B} \sum_{i=1}^{B} \sum_{\ell=1}^{d_i-1} \max\big(0,\ r(t_i^{(\ell)}) - r(t_i^{(\ell+1)}) + m\big), \tag{7}$$

where $t_i^{(\ell)}$ is the label-text embedding of Sec. 3.2 and m is a margin. Penalizing $r(t_i^{(\ell)}) - r(t_i^{(\ell+1)}) + m > 0$ drives finer categories to lie at least a margin farther from the origin than their coarser ancestors, aligning radial position with semantic granularity. The ranking loss acts on label texts only; images are placed relative to them by the entailment loss, which keeps each image inside the cones of its ancestor labels, cones that open away from the origin.

3.4 Training Objective

Our overall training objective is a weighted sum of three terms:

$$L_{\text{total}} = L_{\text{hier}} + \lambda_{\text{ent}}\, L_{\text{entail}} + \lambda_{\text{rank}}\, L_{\text{rank}}. \tag{8}$$

Let L be the maximum taxonomy depth, so that $d_i \le L$. The hierarchical supervised contrastive loss L_{hier} (replacing the original contrastive loss) and the entailment loss are both applied at every level. At level ℓ, the entailment term pairs each annotated image only with the text of its own level-ℓ ancestor, with no negatives and no cross-sample pairs, and averages over these pairs:

$$L_{\text{entail}}^{(\ell)} = \frac{1}{|\mathcal{V}_\ell|} \sum_{i\in\mathcal{V}_\ell} L_{\text{entail}}\big(t_i^{(\ell)},\ z_i\big), \tag{9}$$

where $L_{\text{entail}}(\cdot,\cdot)$ is the per-pair loss of Eq. 5. The per-level terms are aggregated with level weights,

$$L_{\text{hier}} = \sum_{\ell=1}^{L} \lambda_\ell \cdot \tfrac{1}{2}\big(L_\ell^{i\to t} + L_\ell^{t\to i}\big), \qquad L_{\text{entail}} = \sum_{\ell=1}^{L} L_{\text{entail}}^{(\ell)}, \tag{10}$$

Because every per-level term is computed over $\mathcal{V}_\ell$, samples whose path terminates early do not contribute to deeper levels. All loss weights: $\{\lambda_\ell\}$, λ_{ent}, and λ_{rank} are selected by the overall hierarchical precision; Appendix C gives the selection procedure and the chosen values.

4 Experiments and Results

To assess whether HyperRS encodes the label hierarchy beyond per-level accuracy, we evaluate it from three complementary angles: (i) *predictive accuracy*, whether it predicts the correct label at every level; (ii) *structural fidelity*, whether its errors stay within the correct coarse subtree; and (iii) *traversal analysis*, whether the hierarchy can be read off directly by walking through the embedding space. Secs. 4.1–4.3 present the three evaluations in turn, followed by robustness to image corruptions (Sec. 4.4) and ablations (Sec. 4.5).

Datasets. We evaluate on two hierarchical aerial-scene datasets. **Million-AID** (Long et al., 2021) has a three-level taxonomy of mixed depth with 8, 28, and 37 classes at levels 1–3; 14 leaves terminate at level 2 and 37 at level 3, giving 51 leaf classes. We split its 10,000 labeled training images $80\%/20\%$ into 8,001 images for fine-tuning and 1,999 for evaluation. **RSI-CB256** (Li et al., 2020) (hereafter RSI-CB) contains about 24,000 images of 256×256 pixels in a two-level taxonomy with 6 and 35 classes, all leaves lying at level 2; we use 10% of the images for fine-tuning and 90% for evaluation. Both splits are random and stratified by class.

Baselines. We compare with two general-domain image–text models, the Euclidean CLIP ViT-L/16 and the hyperbolic MERU ViT-L/16, both released by Desai et al. (2023) with the same architecture and pretraining data, and with three remote sensing foundation models: the image–text model RemoteCLIP (Liu et al., 2024), and the masked image modeling (MIM) models SatMAE (Cong et al., 2022) and SatMAE++ (Noman et al., 2024). None of these models is pretrained with taxonomic supervision, so we fine-tune each under our multi-level protocol. Each image–text model is trained in three settings: *LeafCon*, the standard contrastive loss on leaf labels only; *HierCon*, the standard contrastive loss applied at every taxonomy level; and *HSC*, our hierarchical supervised contrastive loss (Sec. 3.2). The MIM backbones have no text encoder, so we freeze them and train one linear head per level. All methods are fine-tuned on the same training split of each dataset.

Implementation details. HyperRS is initialized from MERU ViT-L/16 (Desai et al., 2023). We fine-tune it for about 10 epochs at batch size 256 on 224×224 inputs with AdamW (learning rate 1×10^{-5} for both encoders, weight decay 0.2), a linear warmup, and cosine decay, which amounts to 312 iterations with 12 warmup steps on Million-AID and 96 iterations with 5 warmup steps on RSI-CB. The curvature c and temperature τ are learned, and the ranking margin is m=0.1. The loss weights are $\{\lambda_\ell\} = (1, 3, 2.4)$, $\lambda_{\text{ent}} = 3.0$, and $\lambda_{\text{rank}} = 0.5$ on Million-AID, and $\{\lambda_\ell\} = (1, 1)$, $\lambda_{\text{ent}} = 0.2$, and $\lambda_{\text{rank}} = 0.2$ on RSI-CB. All loss weights are selected by hP on the evaluation split, and the level weights are shared by all image–text methods (Appendix C).

Evaluation metrics. We use three groups of metrics (full definitions in Appendix B). The first two evaluate a model's *predictions* (Sec. 4.1); the third evaluates *geometric traversals* of the embedding space (Sec. 4.3). All metrics are reported as percentages.

Prediction metrics capture how accurately a model predicts across the hierarchy: **per-level accuracy** ($\mathrm{Acc}^{(\ell)}$), correctness at each taxonomy level; **hierarchical precision** (hP) (Kiritchenko et al., 2005), the overlap between predicted and ground-truth ancestor sets; **Wu–Palmer similarity** (WuP) (Wu & Palmer, 1994), which scores a prediction by the depth of its lowest common ancestor with ground truth; and the **exact-match ratio** (EMR) (Zhang & Zhou, 2013), the proportion of correct predictions at all levels.

Structural metrics capture whether errors respect the taxonomy. **Conditional Hierarchical Accuracy** (CHA) is evaluated on samples misclassified at the finest level and measures how often the prediction stays correct at coarser levels. Different methods misclassify different images, so we compare HyperRS with each baseline on their *shared* error set $\mathcal{E}_\cap$, the images all compared methods misclassify at the finest level, so that the two are scored on the same samples. We report it under two schemes. In the *independent* scheme, $\mathrm{CHA}^{(\ell)}$ is the fraction of finest-level errors whose level-ℓ prediction is correct; in the *induced* scheme, the level-ℓ label is read off the finest-level prediction by projecting it up the taxonomy, so a coarse level counts as correct only if the fine prediction lies in the right subtree.

Traversal metrics assess whether the hierarchy can be read off geometrically, by walking from an image embedding toward the origin and comparing the retrieved chain of unique nodes against the ground-truth ancestor chain (Sec. 4.3 details the procedure). We measure three aspects of the retrieved chain: **set overlap** with hP, computed between the retrieved and ground-truth node sets; **ordering** with the depth-normalized **longest common subsequence** (LCS) of the two chains; and coverage with a family of **hit rates**. The per-level hit rate $\mathrm{H}_{d\ell}$ records whether the correct level-ℓ ancestor is visited, while two stricter variants require either that *all* ground-truth ancestors be visited ($\mathrm{H}_{\mathrm{all}}$) or that the retrieved nodes match the ground-truth ancestors exactly, with no missing or spurious nodes ($\mathrm{H}_{\mathrm{exact}}$).

Table 1: **Hierarchical classification performance on Million-AID and RSI-CB.** Training settings: Per-level, one linear head per level for backbones without a text encoder; LeafCon, standard image–text contrastive loss on leaf labels only; HierCon, standard contrastive loss at every taxonomy level; HSC, our hierarchical supervised contrastive loss (Sec. 3.2). HyperRS adds the entailment and ranking objectives to HSC on MERU. All values are in %. Best per column in **bold**.

Method	Setting	Million-AID						RSI-CB				
		$\mathrm{Acc}^{(1)}$	$\mathrm{Acc}^{(2)}$	$\mathrm{Acc}^{(3)}$	hP	WuP	EMR	$\mathrm{Acc}^{(1)}$	$\mathrm{Acc}^{(2)}$	hP	WuP	EMR
SatMAE	Per-level	89.6	84.4	80.7	81.2	75.3	68.3	92.2	94.9	93.1	90.7	89.0
SatMAE++	Per-level	90.8	89.5	86.6	84.7	79.6	74.2	92.2	95.4	93.3	90.7	89.2
RemoteCLIP	LeafCon	61.3	70.9	**91.1**	65.6	41.5	36.7	57.9	94.2	78.1	60.9	58.4
	HierCon	92.1	91.0	91.0	84.7	76.2	72.0	91.7	92.6	91.4	88.5	85.4
	HSC (Ours)	92.8	91.4	91.0	85.3	77.7	73.5	92.6	93.3	92.1	89.5	86.8
CLIP	LeafCon	41.6	64.4	89.3	57.8	30.4	24.4	46.0	95.9	70.6	44.9	44.0
	HierCon	92.8	92.1	87.2	85.1	77.6	72.0	95.6	94.4	94.1	92.5	90.0
	HSC (Ours)	93.5	92.9	88.2	85.9	78.7	73.6	95.5	94.6	94.1	92.5	90.0
MERU	LeafCon	45.7	67.0	88.0	56.3	27.9	18.5	48.8	**96.7**	72.3	47.6	46.8
	HierCon	94.2	**93.0**	87.0	86.8	81.2	73.6	95.6	95.4	94.5	92.8	90.9
	HSC (Ours)	**94.6**	92.9	89.0	87.1	81.3	**75.8**	**95.9**	95.7	**94.9**	**93.1**	**91.4**
HyperRS	HSC (Ours)	**94.6**	**93.0**	89.6	**87.2**	**81.8**	**75.8**	**95.9**	95.7	**94.9**	**93.1**	**91.4**

4.1 Superior Predictive Accuracy

Table 1 compares predictive performance under different supervision schemes on Million-AID and RSI-CB. ❶ Per-level linear classifiers attain competitive accuracy at individual levels, but their independently trained predictions yield lower structural scores than the hierarchy-aware supervision models. ❷ Leaf-only contrastive training often achieves high accuracy at the finest level while performing poorly at coarser levels. Applying contrastive supervision at every taxonomy level (HierCon and HSC) substantially improves this balance. ❸ Our HSC objective generally improves on HierCon, particularly in hierarchy-aware metrics, showing the benefit of using shared labels as additional positives. ❹ Building on HSC, HyperRS achieves the strongest hP on both datasets: 87.2% on Million-AID and 94.9% on RSI-CB256. It also improves accuracy at most levels over Euclidean and hyperbolic models trained with the same HSC objective.

4.2 Better Structural Fidelity

Table 2 reports CHA on the shared error set $\mathcal{E}_{\cap}$; a higher CHA means that, when a model errs at the finest level, its prediction more often stays within the correct coarse subtree. As $\mathcal{E}_{\cap}$ varies with the methods being compared, we report CHA pairwise, each block evaluating HyperRS and one baseline on the images both misclassify at the finest level; error sets are thus shared within a block but differ across blocks. HyperRS achieves better CHA scores than all compared baselines on both datasets. Because CHA is measured only on samples already misclassified at the finest level, these gains show that even when it fails at the hardest, fine-grained level, HyperRS preserves correct coarse-level predictions more often than the compared baselines. This graceful degradation suggests that HyperRS is more trustworthy in practice: a mistake still yields a usefully correct coarse label instead of a wholly wrong one.

4.3 Traversal analysis

We examine whether the hierarchy can be read off *geometrically*: moving an image embedding toward the root should pass through progressively more generic concepts, tracing its ancestor chain from fine to coarse. For each evaluation image, we move its embedding toward the root in 50 steps and retrieve the nearest taxonomy node at each step. The root is the origin $\mathbf{O}$ for the hyperbolic models, and the mean of the L_2-normalized training image and text embeddings for the Euclidean ones (Desai et al., 2023). Keeping each retrieved node once, in the order encountered, gives a fine-to-coarse chain that we compare with the ground-truth ancestor chain (Appendix D). The baselines are trained with HierCon.

Table 2: **Conditional Hierarchical Accuracy (CHA) on Million-AID and RSI-CB.** On the two-level RSI-CB, only $\mathrm{CHA}^{(1)}$ is defined, and it coincides with $\mathrm{CHA}_{\mathrm{all}}$. All values are in %. The better value in each pair is in **bold**.

Method	Million-AID					RSI-CB	
	Independent			Induced		Indep.	Induced
	$\mathrm{CHA}^{(1)}$	$\mathrm{CHA}^{(2)}$	$\mathrm{CHA}_{\mathrm{all}}$	$\mathrm{CHA}^{(1)}$	$\mathrm{CHA}^{(2)}$	$\mathrm{CHA}^{(1)}$	$\mathrm{CHA}^{(1)}$
SatMAE	38.7	29.5	20.8	24.7	7.9	38.0	68.0
HyperRS	**43.0**	**33.7**	**23.7**	**46.6**	**27.4**	**39.6**	**70.3**
SatMAE++	45.0	**36.2**	25.0	37.1	16.8	56.0	29.9
HyperRS	**47.5**	35.0	**30.4**	**44.8**	**25.0**	**64.8**	**41.9**
RemoteCLIP	33.3	28.4	28.4	34.2	**21.1**	55.0	30.0
HyperRS	**48.5**	**37.3**	**31.4**	**40.8**	**21.1**	**60.8**	**34.5**
CLIP	51.2	27.2	21.4	38.2	22.2	63.5	55.5
HyperRS	**51.3**	**35.3**	**28.0**	**44.6**	**25.7**	**64.3**	**67.1**
MERU	55.5	39.1	33.3	49.2	28.1	62.6	37.9
HyperRS	**59.2**	**40.2**	**34.5**	**50.0**	**30.8**	**63.4**	**38.4**

Table 3: Traversal-based hierarchical evaluation on Million-AID.

Method	Set overlap	Order	Hit rate				
	hP	LCS	H_{d1}	H_{d2}	H_{d3}	$\mathrm{H}_{\mathrm{all}}$	$\mathrm{H}_{\mathrm{exact}}$
RemoteCLIP	57.9	42.5	13.1	39.9	**85.8**	2.9	2.8
CLIP	57.3	37.7	26.1	31.1	61.4	0.0	0.0
MERU	67.2	58.9	27.1	82.9	82.0	22.6	17.8
HyperRS (ours)	**69.7**	**63.7**	**46.1**	**83.3**	81.8	**36.4**	**23.2**

Table 3 compares how well the embedding geometry of each model recovers the taxonomy on Million-AID. ❶ Euclidean models rarely recover a complete ancestor chain: RemoteCLIP attains the highest finest-level hit rate (H_{d3} 85.8%), yet CLIP and RemoteCLIP recover the whole path for at most 2.9% of the images. ❷ Hyperbolic geometry makes the taxonomy more readable: MERU raises hP from at most 57.9% to 67.2% and $\mathrm{H}_{\mathrm{all}}$ from at most 2.9% to 22.6%. ❸ HyperRS achieves the best score on every traversal metric except H_{d3}; its gain over MERU concentrates at the coarsest level (H_{d1} 46.1% vs. 27.1%) and lifts $\mathrm{H}_{\mathrm{all}}$ and $\mathrm{H}_{\mathrm{exact}}$ to 36.4% and 23.2%. Because these chains come from the embedding geometry alone, the results suggest that the ancestors of an image can be read off by moving its HyperRS embedding toward the origin. Euclidean models essentially lack this property, as their normalized embeddings provide no radial direction along which to order concepts by specificity.

4.4 Robustness to Common Remote Sensing Corruptions

Overhead imagery is routinely degraded by sensor and acquisition artifacts, so a practical remote sensing model should retain its hierarchical predictions under such corruptions. We evaluate robustness on the Million-AID evaluation split, applying three corruptions that are common in remote sensing—missing data regions (*data gaps*), *Gaussian noise*, and *motion blur*—using the corruption simulators of REOBench (Li et al., 2026), and compare HyperRS with the Euclidean CLIP baseline trained with the same HSC objective. Under all three corruptions, HyperRS attains a higher hP than CLIP, and its errors stay closer to the correct path, with the whole-path $\mathrm{CHA}_{\mathrm{all}}$ improving by up to 6.9 points (Table 6 in Appendix E). The largest accuracy gain, 5.7 points, occurs at the finest level under data gaps. The hierarchical advantage of HyperRS therefore persists when the input is degraded.

4.5 Ablation Study

We ablate the two design choices behind HyperRS: supervising all taxonomy levels jointly, and the entailment and ranking objectives added on top of the hyperbolic HSC loss.

Table 4: **Single-level (top) vs. joint (middle and bottom) supervision on Million-AID.** Top: HSC trained on a single level (shaded cells mark the trained level). Middle: HXE (Bertinetto et al., 2020), where α controls its depth-dependent loss weights. Bottom: joint HSC (ours).

Loss	Weights	$\mathrm{Acc}^{(1)}$	$\mathrm{Acc}^{(2)}$	$\mathrm{Acc}^{(3)}$	hP
HSC	$(1, 0, 0)$	**95.0**	34.3	23.5	52.8
	$(0, 1, 0)$	51.3	**93.8**	34.0	55.6
	$(0, 0, 1)$	32.8	37.9	88.6	46.9
HXE	$\alpha = 0.1$	50.5	71.2	88.6	59.5
	$\alpha = 0.5$	54.2	73.5	87.8	61.9
HSC	$(1, 3, 2.4)$	94.6	92.9	**89.0**	**87.1**

Table 5: Ablation of loss components on Million-AID. All variants build on the hyperbolic model with our hierarchical supervised contrastive loss.

	Prediction				Traversal						
	$\mathrm{Acc}^{(1)}$	$\mathrm{Acc}^{(2)}$	$\mathrm{Acc}^{(3)}$	hP	hP	LCS	H_{d1}	H_{d2}	H_{d3}	$\mathrm{H}_{\mathrm{all}}$	$\mathrm{H}_{\mathrm{exact}}$
base (+ HSC)	94.6	92.9	89.0	87.1	67.0	59.6	28.8	80.1	81.4	24.0	17.4
+ entail	94.6	**93.0**	89.4	**87.2**	66.9	59.9	30.3	81.2	80.6	24.9	18.6
+ rank	**94.7**	92.7	88.0	86.8	**70.1**	**63.8**	**46.2**	**84.3**	**82.5**	**37.7**	**23.8**
+ both (Ours)	94.6	**93.0**	**89.6**	**87.2**	69.7	63.7	46.1	83.3	81.8	36.4	23.2

Effect of hierarchy supervision. Table 4 highlights two findings. ❶ Single-level training performs well at the supervised level but poorly at the others, whereas joint HSC achieves strong accuracy across all three levels. Specifically, joint training improves hP from 46.9% with finest-level supervision to 87.1%. We select the joint supervision weights $(1, 3.0, 2.4)$ by grid search on the baseline and reuse them across joint HSC variants (Appendix C). Joint supervision also yields higher CHA than training on the finest level only (Fig. 5 in Appendix F). ❷ Our proposed HSC outperforms HXE, a well-established structured hierarchical classification framework proposed in Bertinetto et al. (2020), supporting its effectiveness over this alternative structured loss.

Loss components. Table 5 ablates the entailment and ranking objectives on the hyperbolic base. The *ranking* objective sharply improves every traversal metric (whole-path $\mathrm{H}_{\mathrm{all}}$ $24.0\% \rightarrow 37.7\%$) but slightly lowers finest-level accuracy ($\mathrm{Acc}^{(3)}$ $89.0\% \rightarrow 88.0\%$). Adding the *entailment* objective complements it: finest-level accuracy returns to its best ($\mathrm{Acc}^{(3)}$ 89.6%) while almost all of the traversal gain is kept ($\mathrm{H}_{\mathrm{all}}$ 36.4%). HyperRS therefore adopts both.

5 Conclusion

We presented HyperRS, a hierarchical fine-tuning framework that aligns a pretrained hyperbolic image–text model with a remote sensing taxonomy. It combines a hierarchical supervised contrastive loss (HSC), which supervises every taxonomy level, with a ranking loss and an entailment loss that order label and image embeddings by their distance from the origin. We evaluated the representation from three angles: prediction accuracy across levels, the structure of errors on misclassified images, and the recovery of ancestor chains by traversing the embedding space. On Million-AID and RSI-CB256, HSC matches or improves hierarchical precision over leaf-only and per-level contrastive training for every backbone, the hyperbolic backbone outperforms its Euclidean counterpart in hierarchical precision under the same supervision, and HyperRS more often keeps the coarse labels correct when its finest prediction fails. The radial constraints further make ancestor chains recoverable from the geometry, which Euclidean embeddings essentially do not support. Even so, HyperRS reaches the coarsest ancestor in fewer than half of the traversals; closing this gap and extending the approach to deeper taxonomies and other earth observation tasks are directions for future work.

AI USE STATEMENT

In this work, we used generative AI tools to help interpret results, by drafting textual descriptions of the experimental results from tables produced by the authors, and to assist with translating the authors' notes into English. We have not used generative AI tools to generate data, develop theoretical models, formulate mathematical claims or proofs, propose hypotheses, design the method or experiments, or implement the method, and the remaining required disclosure tasks are not applicable to this work. Additionally, we used generative AI tools to draft and edit parts of the paper for readability and to identify relevant literature. We have reviewed all AI-assisted work: every reported number comes from our own experiments and was checked against the text, and every citation was verified against its source. We take responsibility for the final content of this work, including text, claims, or artifacts produced with the aid of generative AI.

ETHICS STATEMENT

This work uses only publicly released remote sensing benchmarks, Million-AID (Long et al., 2021) and RSI-CB256 (Li et al., 2020), together with publicly released pretrained models, and involves no human subjects or personal data. The images are scene-level aerial and satellite photographs labeled with land-use categories. As with other scene classification methods, improved recognition of land-use categories could be applied to surveillance; our contribution concerns representation learning and evaluation and introduces no capability specific to such uses.

REPRODUCIBILITY STATEMENT

The datasets, splits, baselines, and training settings, including all loss weights, are described in Sec. 4. Appendix A gives the full entailment formulation, Appendix B defines every evaluation metric, Appendix C describes the loss-weight search, Appendix D details the traversal procedure, and Appendix E specifies the corruption protocol. All backbones are initialized from publicly available checkpoints. Our code is provided as anonymized supplementary material and will be released publicly upon publication.

REFERENCES

Mina Ghadimi Atigh, Julian Schoep, Erman Acar, Nanne Van Noord, and Pascal Mettes. Hyperbolic image segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 4453–4462, 2022.

Luca Bertinetto, Romain Mueller, Konstantinos Tertikas, Sina Samangooei, and Nicholas A Lord. Making better mistakes: Leveraging class hierarchies with deep networks. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 12503–12512. IEEE, 2020.

Yezhen Cong, Samar Khanna, Chenlin Meng, Patrick Liu, Erik Rozi, Yutong He, Marshall Burke, David Lobell, and Stefano Ermon. Satmae: Pre-training transformers for temporal and multi-spectral satellite imagery. *Advances in Neural Information Processing Systems*, 35:197–211, 2022.

Karan Desai, Maximilian Nickel, Tanmay Rajpurohit, Justin Johnson, and Shanmukha Ramakrishna Vedantam. Hyperbolic image-text representations. In *International Conference on Machine Learning*, pp. 7694–7731. PMLR, 2023.

Aleksandr Ermolov, Leyla Mirvakhabova, Valentin Khrulkov, Nicu Sebe, and Ivan Oseledets. Hyperbolic vision transformers: Combining improvements in metric learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 7409–7419, 2022.

Octavian Ganea, Gary Bécigneul, and Thomas Hofmann. Hyperbolic entailment cones for learning hierarchical embeddings. In *International conference on machine learning*, pp. 1646–1655. PMLR, 2018.

Manal Hamzaoui, Laetitia Chapel, Minh-Tan Pham, and Sébastien Lefèvre. A hierarchical prototypical network for few-shot remote sensing scene classification. In *International Conference on Pattern Recognition and Artificial Intelligence*, pp. 208–220. Springer, 2022.

Chao Jia, Yinfei Yang, Ye Xia, Yi-Ting Chen, Zarana Parekh, Hieu Pham, Quoc Le, Yun-Hsuan Sung, Zhen Li, and Tom Duerig. Scaling up visual and vision-language representation learning with noisy text supervision. In *International conference on machine learning*, pp. 4904–4916. PMLR, 2021.

Prannay Khosla, Piotr Teterwak, Chen Wang, Aaron Sarna, Yonglong Tian, Phillip Isola, Aaron Maschinot, Ce Liu, and Dilip Krishnan. Supervised contrastive learning. *Advances in neural information processing systems*, 33:18661–18673, 2020.

Valentin Khrulkov, Leyla Mirvakhabova, Evgeniya Ustinova, Ivan Oseledets, and Victor Lempitsky. Hyperbolic image embeddings. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 6418–6428, 2020.

Svetlana Kiritchenko, Stan Matwin, A Fazel Famili, et al. Functional annotation of genes using hierarchical text categorization. In *Proc. of the ACL Workshop on Linking Biological Literature, Ontologies and Databases: Mining Biological Semantics*, volume 76, 2005.

Svetlana Kiritchenko, Stan Matwin, Richard Nock, and A Fazel Famili. Learning and evaluation in the presence of class hierarchies: Application to text categorization. In *Conference of the Canadian society for computational studies of intelligence*, pp. 395–406. Springer, 2006.

Aris Kosmopoulos, Ioannis Partalas, Eric Gaussier, Georgios Paliouras, and Ion Androutsopoulos. Evaluation measures for hierarchical classification: a unified view and novel approaches. *Data Mining and Knowledge Discovery*, 29(3):820–865, 2015.

John M Lee. *Introduction to Riemannian manifolds*, volume 2. Springer, 2018.

Haifeng Li, Xin Dou, Chao Tao, Zhixiang Wu, Jie Chen, Jian Peng, Min Deng, and Ling Zhao. Rsi-cb: A large-scale remote sensing image classification benchmark using crowdsourced data. *Sensors*, 20(6):1594, 2020.

Xiang Li, Yong Tao, Siyuan Zhang, Siwei Liu, Zhitong Xiong, Chunbo Luo, Lu Liu, Mykola Pechenizkiy, Xiaoxiang Zhu, and Tianjin Huang. Reobench: Benchmarking robustness of earth observation foundation models. *Advances in Neural Information Processing Systems*, 38, 2026.

Fan Liu, Delong Chen, Zhangqingyun Guan, Xiaocong Zhou, Jiale Zhu, Qiaolin Ye, Liyong Fu, and Jun Zhou. Remoteclip: A vision language foundation model for remote sensing. *IEEE Transactions on Geoscience and Remote Sensing*, 62:1–16, 2024.

Yang Long, Gui-Song Xia, Shengyang Li, Wen Yang, Michael Ying Yang, Xiao Xiang Zhu, Liangpei Zhang, and Deren Li. On creating benchmark dataset for aerial image interpretation: Reviews, guidances, and million-aid. *IEEE Journal of selected topics in applied earth observations and remote sensing*, 14:4205–4230, 2021.

Maximillian Nickel and Douwe Kiela. Poincaré embeddings for learning hierarchical representations. *Advances in neural information processing systems*, 30, 2017.

Maximillian Nickel and Douwe Kiela. Learning continuous hierarchies in the lorentz model of hyperbolic geometry. In *International conference on machine learning*, pp. 3779–3788. PMLR, 2018.

Mubashir Noman, Muzammal Naseer, Hisham Cholakkal, Rao Muhammad Anwer, Salman Khan, and Fahad Shahbaz Khan. Rethinking transformers pre-training for multi-spectral satellite imagery. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 27811–27819, 2024.

Avik Pal, Max Van Spengler, Guido D'Amely di Melendugno, Alessandro Flaborea, Fabio Galasso, and Pascal Mettes. Compositional entailment learning for hyperbolic vision-language models. In *International Conference on Learning Representations*, volume 2025, pp. 87371–87399, 2025.

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*, pp. 8748–8763. PMLR, 2021.

Colorado J Reed, Ritwik Gupta, Shufan Li, Sarah Brockman, Christopher Funk, Brian Clipp, Kurt Keutzer, Salvatore Candido, Matt Uyttendaele, and Trevor Darrell. Scale-mae: A scale-aware masked autoencoder for multiscale geospatial representation learning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 4088–4099, 2023.

Ozlem Sen and Hacer Yalim Keles. A hierarchical approach to remote sensing scene classification. *PFG–Journal of Photogrammetry, Remote Sensing and Geoinformation Science*, 90(2):161–175, 2022.

Aditya Sinha, Siqi Zeng, Makoto Yamada, and Han Zhao. Learning structured representations with hyperbolic embeddings. *Advances in Neural Information Processing Systems*, 37:91220–91259, 2024.

Zhibiao Wu and Martha Palmer. Verb semantics and lexical selection. In *32nd annual meeting of the association for computational linguistics*, pp. 133–138, 1994.

Peng Zeng, Shixuan Lin, Hao Sun, and Dongbo Zhou. Exploiting hierarchical label information in an attention-embedding, multi-task, multi-grained, network for scene classification of remote sensing imagery. *Applied Sciences*, 12(17):8705, 2022.

Min-Ling Zhang and Zhi-Hua Zhou. A review on multi-label learning algorithms. *IEEE transactions on knowledge and data engineering*, 26(8):1819–1837, 2013.

Shu Zhang, Ran Xu, Caiming Xiong, and Chetan Ramaiah. Use all the labels: A hierarchical multi-label contrastive learning framework. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 16639–16648. IEEE, 2022.

Zilun Zhang, Tiancheng Zhao, Yulong Guo, and Jianwei Yin. Rs5m and georsclip: A large-scale vision-language dataset and a large vision-language model for remote sensing. *IEEE Transactions on Geoscience and Remote Sensing*, 62:1–23, 2024.

A Entailment Cone Details

This appendix gives the full definition of the exterior angle $\mathrm{ext}(\cdot,\cdot)$ used in the entailment loss of Sec. 3.1 (Eqs. 4–5), following MERU (Desai et al., 2023). We work in the Lorentz model of n-dimensional hyperbolic geometry with curvature $-c$ ($c > 0$); recall from Sec. 3.1 that a point $\mathbf{x} = [x_{\text{time}}, \mathbf{x}_{\text{space}}]$ lies on the hyperboloid $\mathcal{L}^n$ with $\langle \mathbf{x}, \mathbf{x} \rangle_{\mathcal{L}} = -1/c$, that the origin is $\mathbf{O} = [\,1/\sqrt{c},\, \mathbf{0}\,]$, and that $\langle \cdot, \cdot \rangle_{\mathcal{L}}$ denotes the Lorentzian inner product (Eq. 1).

Entailment cones. For a point $\mathbf{x} \in \mathcal{L}^n$ (other than the origin), Following Ganea et al. (2018) and MERU, each point defines a cone with apex at $\mathbf{x}$, opening along the geodesic that connects the origin $\mathbf{O}$ to $\mathbf{x}$ and extending outward (away from the origin). A point $\mathbf{y}$ is said to be *entailed* by $\mathbf{x}$ when it lies inside this cone. The cone's half-aperture is

$$\mathrm{aper}(\mathbf{x}) = \sin^{-1}\left(\min\left\{ \frac{2K}{\sqrt{c}\,\|\mathbf{x}_{\text{space}}\|},\; 1 - \epsilon \right\} \right), \tag{11}$$

with a constant $K = 0.1$ and $\epsilon = 10^{-8}$, as in Eq. 4. Because $\mathrm{aper}(\mathbf{x})$ decreases as $\|\mathbf{x}_{\text{space}}\|$ (and hence the distance from the origin) grows, cones near the origin are wide and cones far from it are narrow: a generic concept near the origin can entail many more specific concepts, while a specific concept far from the origin entails few. This realizes the partial order *text-entails-image*, in which more general nodes sit closer to the origin.

Exterior angle. For an ordered pair $(\mathbf{x}, \mathbf{y})$, the exterior angle at x is the angle between the geodesic ray from x to y and the outward continuation of the geodesic Ox beyond x (equivalently, $\pi - \angle Oxy$, the supplement of the interior angle at x); it measures how far $\mathbf{y}$ deviates from the cone's axis as seen from $\mathbf{x}$. In the Lorentz model it is given in closed form by

$$\mathrm{ext}(\mathbf{x}, \mathbf{y}) = \cos^{-1}\left(\frac{y_{\text{time}} + x_{\text{time}}\left(c\,\langle \mathbf{x}, \mathbf{y} \rangle_{\mathcal{L}}\right)}{\|\mathbf{x}_{\text{space}}\| \sqrt{\left(c\,\langle \mathbf{x}, \mathbf{y} \rangle_{\mathcal{L}}\right)^2 - 1}} \right). \tag{12}$$

When $\mathrm{ext}(\mathbf{x}, \mathbf{y}) \leq \mathrm{aper}(\mathbf{x})$, the point $\mathbf{y}$ lies inside the cone of $\mathbf{x}$ and the partial order is satisfied; otherwise it is violated.

Entailment loss. The entailment loss penalizes only violations, i.e. the amount by which the exterior angle exceeds the half-aperture,

$$L_{\text{entail}}(\mathbf{x}, \mathbf{y}) = \max\big(0,\; \mathrm{ext}(\mathbf{x}, \mathbf{y}) - \mathrm{aper}(\mathbf{x})\big), \tag{13}$$

which is zero when $\mathbf{y}$ is already entailed by $\mathbf{x}$ and grows with the degree of violation. HyperRS keeps MERU's per-pair formulation with $\mathbf{x}$ the text embedding and $\mathbf{y}$ the image embedding. Whereas MERU pairs each image with its caption, HyperRS pairs each image with the label text of its ancestor at every annotated level and averages the per-level terms as in Eq. 9.

B Evaluation Metrics

This appendix gives the full definitions of the prediction and structural metrics summarized in Sec. 4: per-level accuracy, hierarchical precision (hP), Conditional Hierarchical Accuracy (CHA), Wu–Palmer similarity (WuP), and the exact match ratio (EMR). Their results are reported in Secs. 4.1 and 4.2.

B.1 Notation and Setup

Let $\mathcal{T}$ be a class taxonomy with L granularity levels, rooted at a virtual node that is excluded from all computations. For a node u, let $\mathcal{A}(u)$ denote its *augmented ancestor set*: the set containing u together with all of its ancestors up to, but excluding, the root (Kiritchenko et al., 2005; 2006); and let $\mathrm{anc}_\ell(u)$ denote its unique ancestor at level ℓ. The test set is $\{(x_i, y_i)\}_{i=1}^N$, where y_i is the ground-truth leaf label.

At every level $\ell \in \{1, \dots, L\}$ we form the candidate set $\mathcal{C}_\ell$ of level-ℓ classes and a per-level prediction $\hat{y}_i^{(\ell)}$ (produced as described in the Protocol of Sec. 4.1). The level-ℓ ground truth is $y_i^{(\ell)} = \mathcal{A}(y_i) \cap \mathcal{C}_\ell$; since $\mathcal{T}$ is a tree, this intersection contains exactly one class whenever it is non-empty. Because the taxonomy has mixed depth, some leaf classes terminate before the deepest level and therefore have no labels at subsequent levels; we write $\mathcal{V}_\ell = \{\, i : y_i^{(\ell)} \text{ exists} \,\}$ for the samples *valid* at level ℓ, d_i for the deepest level at which y_i is annotated, and $\hat{y}_i \equiv \hat{y}_i^{(d_i)}$ for the finest-level prediction. Finally, $\mathcal{S}_c = \{\, i : y_i = c \,\}$ denotes the samples whose ground-truth leaf is c, and $\mathcal{C}_{leaf}$ the set of leaf classes with $\mathcal{S}_c \neq \emptyset$, which may lie at different levels; leaf-macro averages first average over $\mathcal{S}_c$ and then over $\mathcal{C}_{leaf}$.

B.2 Per-Level Accuracy

Level-ℓ accuracy measures flat classification quality at a single granularity. To weight every category equally regardless of its sample count, we report the macro-averaged accuracy, which first computes per-class accuracy and then averages over all level-ℓ classes:

$$\mathrm{Acc}^{(\ell)} = \frac{1}{|\mathcal{C}_\ell|} \sum_{c \in \mathcal{C}_\ell} \frac{1}{|\mathcal{V}_\ell^c|} \sum_{i \in \mathcal{V}_\ell^c} \mathbb{1}\left[\hat{y}_i^{(\ell)} = y_i^{(\ell)}\right], \tag{14}$$

where $\mathcal{V}_\ell^c = \{\, i \in \mathcal{V}_\ell : y_i^{(\ell)} = c \,\}$ is the set of level-ℓ samples whose ground-truth class is c. Samples not valid at level ℓ are excluded from its denominator.

B.3 Hierarchical Precision

To summarize performance across all levels into a single number, we use hierarchical precision (Kiritchenko et al., 2005; 2006), a standard set-based metric. For sample i, let $\mathcal{T}_i = \mathcal{A}(y_i)$ be the ground-truth ancestor set and $\mathcal{P}_i = \{\, \hat{y}_i^{(\ell)} : \ell \in \{1, \dots, L\},\ i \in \mathcal{V}_\ell \,\}$ its per-level predictions. The per-sample hierarchical precision is the overlap

$$\mathrm{hP}_i = \frac{|\mathcal{P}_i \cap \mathcal{T}_i|}{|\mathcal{P}_i|}. \tag{15}$$

Because each leaf class corresponds to a unique root-to-leaf path, we report a *leaf-macro* average that weights every path equally. Let $\mathcal{C}_{leaf}$ be the set of leaf classes and $\mathcal{S}_c = \{\, i : y_i = c \,\}$ the samples with ground-truth leaf c; then

$$\mathrm{hP} = \frac{1}{|\mathcal{C}_{leaf}|} \sum_{c \in \mathcal{C}_{leaf}} \frac{1}{|\mathcal{S}_c|} \sum_{i \in \mathcal{S}_c} \mathrm{hP}_i. \tag{16}$$

B.4 Conditional Hierarchical Accuracy (CHA)

CHA probes graceful degradation: on samples misclassified at the finest level, it measures how often the coarser ancestors of the prediction remain correct. We call a model *structurally correct* for sample i at level ℓ when its level-ℓ prediction matches the ground truth, $\hat{y}_i^{(\ell)} = y_i^{(\ell)}$, and the model *errs* on sample i when it is structurally incorrect at that sample's own deepest level. The per-level predictions $\hat{y}_i^{(\ell)}$ are obtained under one of two schemes. In the *induced* scheme, the model predicts only at the finest level and its coarser predictions are derived through the taxonomy, $\hat{y}_i^{(\ell)} = \mathrm{anc}_\ell(\hat{y}_i)$, so all levels reflect a single decision. In the *independent* scheme, a separate zero-shot classifier predicts $\hat{y}_i^{(\ell)}$ at each level, and predictions across levels need not be mutually consistent.

Let $\mathcal{E}$ be the shared set of finest-level errors of the two compared methods, and let $\mathcal{E}_c = \{i \in \mathcal{E} : y_i = c\}$ for each ground-truth deepest-level class $c \in \mathcal{C}_\mathrm{L}$. For level ℓ, define $\mathcal{E}_{\ell,c} = \{i \in \mathcal{E}_c : \ell < d_i\}$. We average only over deepest-level classes with eligible errors: $\mathcal{C}_\ell^{\mathcal{E}} = \{c \in \mathcal{C}_\mathrm{L} : |\mathcal{E}_{\ell,c}| > 0\}$ and

$\mathcal{C}^{\mathcal{E}} = \{c \in \mathcal{C}_{\mathrm{L}} : |\mathcal{E}_c| > 0\}$. The leaf-macro-averaged per-level and whole-path CHA are

$$\mathrm{CHA}^{(\ell)} = \frac{1}{|\mathcal{C}_\ell^{\mathcal{E}}|} \sum_{c \in \mathcal{C}_\ell^{\mathcal{E}}} \frac{1}{|\mathcal{E}_{\ell,c}|} \sum_{i \in \mathcal{E}_{\ell,c}} \mathbb{1}\left[\hat{y}_i^{(\ell)} = y_i^{(\ell)}\right], \tag{17}$$

$$\mathrm{CHA}_{\mathrm{all}} = \frac{1}{|\mathcal{C}^{\mathcal{E}}|} \sum_{c \in \mathcal{C}^{\mathcal{E}}} \frac{1}{|\mathcal{E}_c|} \sum_{i \in \mathcal{E}_c} \mathbb{1}\left[\forall\, \ell < d_i,\ \hat{y}_i^{(\ell)} = y_i^{(\ell)}\right]. \tag{18}$$

Level ℓ is scored for sample i only when it is a *strict* ancestor of that sample's leaf ($\ell < d_i$); otherwise the sample's leaf mistake would be counted a second time as an ancestor failure. Thus $\mathrm{CHA}^{(\ell)}$ localizes, level by level, where the coarse context survives an error, while $\mathrm{CHA}_{\mathrm{all}}$ requires the entire surviving path to be correct—i.e. the prediction is wrong only at the deepest level. All rates are macro-averaged over ground-truth classes. Under the induced scheme, coarser predictions are induced through the taxonomy, a correct ancestor at level ℓ implies a correct ancestor at every coarser level, so being structurally correct at the coarsest ancestor level already guarantees the whole path. With a three-level taxonomy this makes $\mathrm{CHA}^{(2)} = \mathrm{CHA}_{\mathrm{all}}$, and we therefore omit the induced $\mathrm{CHA}_{\mathrm{all}}$ column in Table 2.

Shared error set. Comparing each model on its own errors is not a fair comparison: a more accurate model has a smaller error set, and what remains is the residue of intrinsically harder, more ambiguous images—exactly those on which any model recovers least. A naive comparison would therefore penalize the stronger model. Writing $\mathcal{E}^{(m)}$ for the finest-level error set of method m, we compare HyperRS with each baseline b on their shared intersection

$$\mathcal{E}_\cap = \mathcal{E}^{(HyperRS)} \cap \mathcal{E}^{(b)} \tag{19}$$

which is symmetric in the two methods and gives both an identical denominator, isolating graceful degradation from accuracy: no method is credited for being correct, only for failing in a more structured way. Because $\mathcal{E}_\cap$ depends on the baseline, each pair in Table 2 is computed on its own shared error set.

B.5 Wu–Palmer Similarity

Wu–Palmer similarity (Wu & Palmer, 1994) scores how close two nodes are on the taxonomy by the depth of their lowest common ancestor (LCA). We apply it to a top-down prediction $\hat{y}_i^{\mathrm{td}}$, obtained by selecting the highest-scoring level-1 class and then, at each subsequent level, the highest-scoring child of the previous prediction until a leaf is reached. For ground-truth leaf y_i,

$$\mathrm{WuP}_i = \frac{2\,\mathrm{depth}\big(\mathrm{LCA}(\hat{y}_i^{\mathrm{td}}, y_i)\big)}{\mathrm{depth}(\hat{y}_i^{\mathrm{td}}) + \mathrm{depth}(y_i)}, \tag{20}$$

where $\mathrm{depth}(\cdot)$ is the number of edges from the root and $\mathrm{LCA}(\cdot,\cdot)$ is the lowest common ancestor. $\mathrm{WuP}_i \in [0,1]$ equals 1 for a correct prediction and decreases as the prediction and ground truth diverge higher up the tree, so it rewards near misses (a sibling of the true class) over distant confusions. As for hP, we report the leaf-macro average over ground-truth leaf classes,

$$\mathrm{WuP} = \frac{1}{|\mathcal{C}_{leaf}|} \sum_{c \in \mathcal{C}_{leaf}} \frac{1}{|\mathcal{S}_c|} \sum_{i \in \mathcal{S}_c} \mathrm{WuP}_i. \tag{21}$$

B.6 Exact Match Ratio

The exact match ratio (EMR) is the fraction of samples whose predictions are correct at every evaluated level simultaneously:

$$\mathrm{EMR} = \frac{1}{|\mathcal{C}_{leaf}|} \sum_{c \in \mathcal{C}_{leaf}} \frac{1}{|\mathcal{S}_c|} \sum_{i \in \mathcal{S}_c} \mathbb{1}\left[\forall\, \ell \le d_i :\ \hat{y}_i^{(\ell)} = y_i^{(\ell)}\right], \tag{22}$$

leaf-macro averaged over ground-truth classes. Unlike CHA, which conditions on finest-level errors, EMR is computed over all samples and is the strictest of our metrics: it credits a prediction only when the whole root-to-leaf path is recovered.

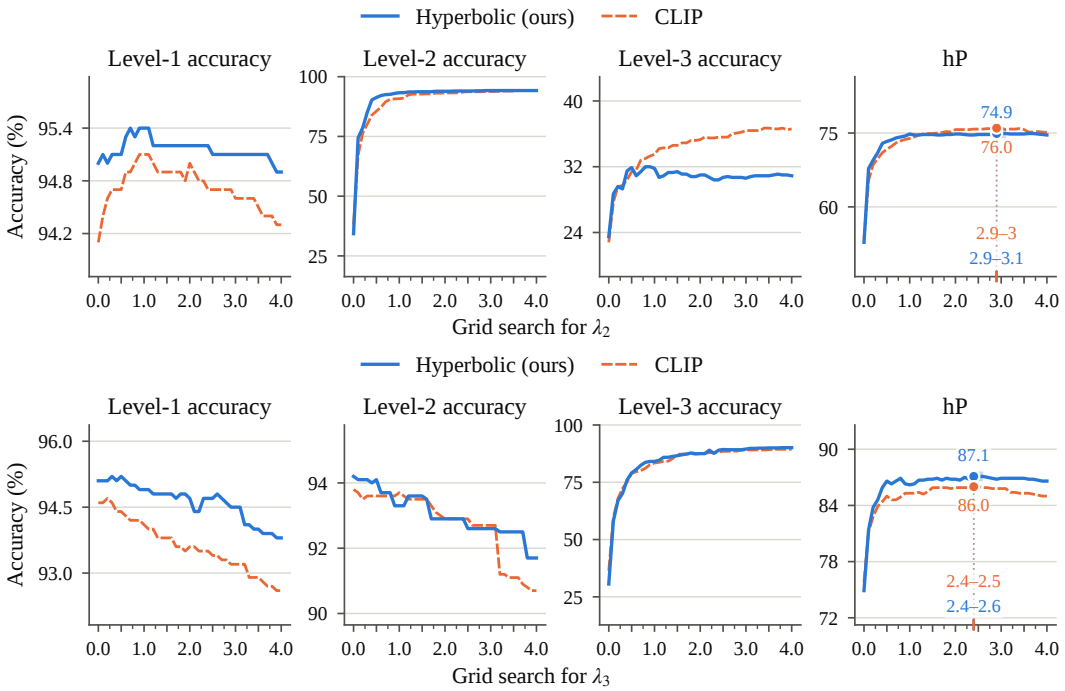

Figure 2: Level-weight search on Million-AID. Hierarchical precision (hP) as a function of the level weights. (a) Sweeping λ_2 with λ_1=1 and λ_3=0. (b) Sweeping λ_3 with λ_1=1 and λ_2=3.0.

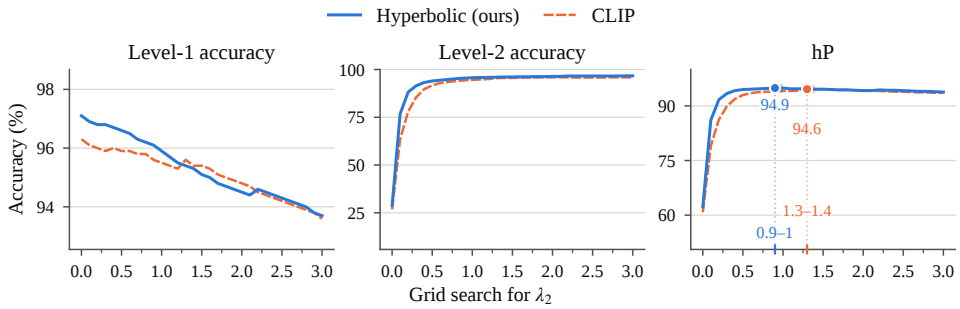

Figure 3: Level-weight search on RSI-CB. Hierarchical precision (hP) for different values of λ_2, with λ_1 fixed to 1.

C Loss-Weight Search

This appendix describes how the loss weights of Eqs. 8 and 10 are selected. Neither dataset has a separate validation split, so every weight is chosen by grid search on the overall hierarchical precision hP of the evaluation split. The search has two stages, applied to each dataset in the same way. First, the level weights $\{\lambda_\ell\}$ of HSC are selected on MERU trained with HSC alone and then shared by every image–text method, so the level weighting is not tuned for any single method. This search proceeds one level at a time with λ_1=1: on Million-AID, we sweep λ_2 with λ_3=0 and then λ_3 with λ_2 fixed (Fig. 2); on the two-level RSI-CB, a single sweep over λ_2 suffices (Fig. 3). Second, with $\{\lambda_\ell\}$ fixed, we select the objective weights λ_{ent} and λ_{rank} on HyperRS; these weights belong to losses that only HyperRS uses, so the baselines have no counterpart to tune. The selected values are $\{\lambda_\ell\} = (1, 3, 2.4)$, $\lambda_{\text{ent}} = 3.0$, and $\lambda_{\text{rank}} = 0.5$ on Million-AID, and $\{\lambda_\ell\} = (1, 1)$, $\lambda_{\text{ent}} = 0.2$, and $\lambda_{\text{rank}} = 0.2$ on RSI-CB.

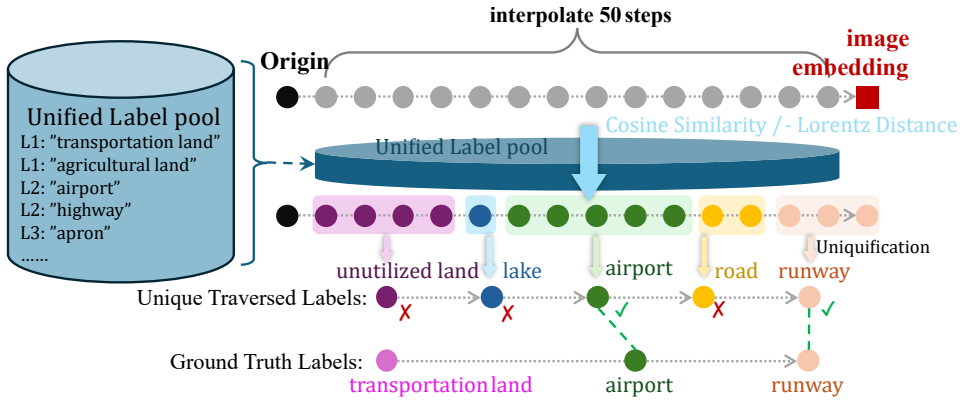

Figure 4: Overview of the traversal procedure.

D TRAVERSAL METRICS

This appendix gives the full traversal procedure and metric definitions summarized in Sec. 4.3. We use the notation of Appendix B: a taxonomy with L levels, augmented ancestor sets $\mathcal{A}(\cdot)$, leaf classes $\mathcal{C}_{leaf}$, the samples $\mathcal{S}_c = \{\, i : y_i = c \,\}$ of leaf class c, ground-truth level-ℓ ancestor $y_i^{(\ell)}$, and deepest annotated level d_i.

Traversal procedure. Figure 4 illustrates the procedure. For every image we interpolate 50 equally spaced steps along the geodesic connecting the image embedding to the origin. For the hyperbolic models we interpolate in the tangent space at the origin and lift each step back onto the hyperboloid; for CLIP we linearly interpolate between the L_2-normalized image embedding and an `[ROOT]` by averaging all training image and text embeddings after L2 normalization. At each step we retrieve the nearest neighbour from the candidate set $\mathcal{X}$ of taxonomy label names—using Lorentzian distance for the hyperbolic models and cosine similarity for CLIP, matching each model's training objective—so that every retrieved item is a taxonomy node. Consecutive steps often return the same label, so we keep only unique labels in the order encountered, giving a retrieved chain $\hat{\pi}_i$ ordered from fine to coarse. We write $\hat{\Pi}_i = \{\, t \in \hat{\pi}_i : t \neq \texttt{[ROOT]} \,\}$ for the retrieved node set with the root removed, and compare against the ground-truth ancestor chain $\mathcal{T}_i = \mathcal{A}(y_i)$.

Set overlap. We report the hierarchical precision of Appendix B, measuring the overlap between $\hat{\Pi}_i$ and $\mathcal{T}_i$, leaf-macro averaged. Because a traversal may visit any number of nodes, visiting labels outside $\mathcal{T}_i$ lowers the score.

Ordering. The longest common subsequence (LCS) between the retrieved and ground-truth chains, both ordered leaf-to-root, normalized by depth $d_i = |\mathcal{T}_i|$:

$$\mathrm{LCS} = \frac{1}{|\mathcal{C}_{leaf}|} \sum_{c \in \mathcal{C}_{leaf}} \frac{1}{|\mathcal{S}_c|} \sum_{i \in \mathcal{S}_c} \frac{|\mathrm{lcs}(\hat{\pi}_i, \mathcal{T}_i)|}{d_i}. \tag{23}$$

Hit rates. At level ℓ, $\mathrm{H}_{d\ell}$ measures whether the correct level-ℓ ancestor was visited. Let $\mathcal{S}_c^{\ell} = \{\, i \in \mathcal{S}_c : \ell \leq d_i \,\}$ be the samples of leaf class c carrying a level-ℓ label, and $\mathcal{C}_{leaf}^{\ell} = \{\, c \in \mathcal{C}_{leaf} : \mathcal{S}_c^{\ell} \neq \emptyset \,\}$ the leaf classes for which such samples exist. Because the taxonomy has mixed depth, a leaf terminating above level ℓ has $\mathcal{S}_c^{\ell} = \emptyset$; such classes are excluded from the level-ℓ average, so that

$$\mathrm{H}_{d\ell} = \frac{1}{|\mathcal{C}_{leaf}^{\ell}|} \sum_{c \in \mathcal{C}_{leaf}^{\ell}} \frac{1}{|\mathcal{S}_c^{\ell}|} \sum_{i \in \mathcal{S}_c^{\ell}} \mathbb{1}\left[y_i^{(\ell)} \in \hat{\Pi}_i \right], \tag{24}$$

Table 6: Robustness on the Million-AID evaluation split under three common remote sensing corruptions. Both models are trained with HSC on clean images; results are averaged over five severity levels. All values are in %. Best per corruption in **bold**.

		Accuracy				CHA(independent)			CHA(induced)	
Corruption	Method	$\mathrm{Acc}^{(1)}$	$\mathrm{Acc}^{(2)}$	$\mathrm{Acc}^{(3)}$	hP	$\mathrm{CHA}^{(1)}$	$\mathrm{CHA}^{(2)}$	$\mathrm{CHA}_{\mathrm{all}}$	$\mathrm{CHA}^{(1)}$	$\mathrm{CHA}^{(2)}$
Data gaps	CLIP	83.6	**74.2**	69.2	71.8	56.3	32.9	30.1	54.7	23.8
	HyperRS	**87.3**	73.9	**74.9**	**74.7**	**56.6**	**40.2**	**33.2**	**55.7**	**23.8**
Gaussian noise	CLIP	91.5	84.8	82.9	80.5	**53.9**	29.9	22.7	45.3	23.2
	HyperRS	**92.9**	**87.0**	**84.0**	**83.4**	53.5	**35.6**	**29.6**	**50.9**	**24.3**
Motion blur	CLIP	90.3	**86.1**	81.9	80.6	49.3	27.0	21.5	**48.5**	24.3
	HyperRS	**90.7**	85.8	**83.9**	**81.1**	**54.2**	**34.5**	**28.2**	46.4	**26.1**

which is well defined since every $c \in \mathcal{C}_{leaf}^{\ell}$ has $|\mathcal{S}_c^{\ell}| \geq 1$. Two stricter variants require the whole path:

$$\mathrm{H}_{\mathrm{all}} = \frac{1}{|\mathcal{C}_{leaf}|} \sum_{c \in \mathcal{C}_{leaf}} \frac{1}{|\mathcal{S}_c|} \sum_{i \in \mathcal{S}_c} \mathbb{1}\left[\mathcal{T}_i \subseteq \hat{\Pi}_i\right], \qquad \mathrm{H}_{\mathrm{exact}} = \frac{1}{|\mathcal{C}_{leaf}|} \sum_{c \in \mathcal{C}_{leaf}} \frac{1}{|\mathcal{S}_c|} \sum_{i \in \mathcal{S}_c} \mathbb{1}\left[\hat{\Pi}_i = \mathcal{T}_i\right]. \tag{25}$$

E Robustness Results

This appendix reports the full results of the corruption study in Sec. 4.4. We compare HyperRS with the Euclidean CLIP ViT-L/16 baseline. Both models are fine-tuned on the clean Million-AID training images with HSC under the same settings (Sec. 4), so they differ only in the embedding geometry and in the entailment and ranking objectives of HyperRS. At test time, we apply three corruptions from REOBench (Li et al., 2026) to the Million-AID evaluation images: *data gaps*, which mask out regions of the image to mimic missing sensor data; *Gaussian noise*; and *motion blur*. Each corruption is applied at the five severity levels of REOBench, and we report the average over them. No corrupted images are used for training.

Table 6 reports per-level accuracy, hP, and CHA under both inference schemes. As in Table 2, CHA is computed on the images that both models misclassify at the finest level under the given corruption.

F Effect of Joint Supervision on CHA

Figure 5 compares CHA between HSC trained on the finest level only, $(0, 0, 1)$, and joint HSC, $(1, 3, 2.4)$, on the images that both models misclassify at the finest level.

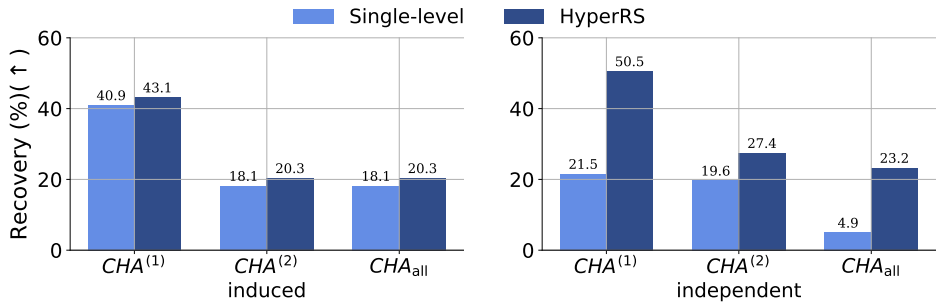

Figure 5: **Effect of joint supervision on CHA.** Finest-level HSC, $(0, 0, 1)$, vs. joint HSC, $(1, 3, 2.4)$, under the induced (left) and independent (right) inference schemes. Joint supervision recovers correct ancestors more often, with the largest gain on the whole-path rate.