ICCV
#9

ICCV
#9

ICCV 2025 Submission #9. CONFIDENTIAL REVIEW COPY. DO NOT DISTRIBUTE.

Hyperbolic Multimodal Representation Learning for Biological Taxonomies

Anonymous ICCV submission

Paper ID 9

Abstract

Taxonomic classification in biodiversity research involves organizing biological specimens into structured hierarchies based on evidence, which can come from multiple modalities such as images and genetic information. We investigated whether hyperbolic networks provide a better embedding space for such hierarchical models. Our method embeds multimodal inputs into a shared hyperbolic space using contrastive and novel entailment-based objectives. Experiments on the BIOSCAN-1M dataset show that hyperbolic embeddings achieve competitive performance with Euclidean baselines, and outperforms all other models on unseen species classification using DNA barcodes. However, fine-grained classification and open-world generalization remain challenging. This framework offers a scalable and structure-aware foundation for biodiversity modelling, with potential applications to species discovery, ecological monitoring, and conservation efforts.

1. Introduction

Taxonomic classification is essential for monitoring and mitigating biodiversity loss, requiring accurate identification of specimens across diverse ecosystems. DNA barcodes [1, 7] provide a way to classify specimens to known taxa or identify them as novel to science, but classification to the species level remains challenging when barcodes are unavailable. To tackle this, Gong et al. [5] showed that using contrastive learning to align DNA barcode embeddings to image embeddings can improve classification at the species level even when only using images as input at inference.

However, a key limitation of CLIBD [5] and other recent biodiversity-focused multimodal methods [17] is that the methods do not utilize the known taxonomic hierarchy of the input data. To address this, we explore whether embeddings in hyperbolic space can better capture the hierarchical structure of taxonomic relationships, enabling better fine-grained classification. While training, the model takes inputs from multiple modalities—DNA barcodes, specimen images, and hierarchical taxonomic labels—and co-aligns
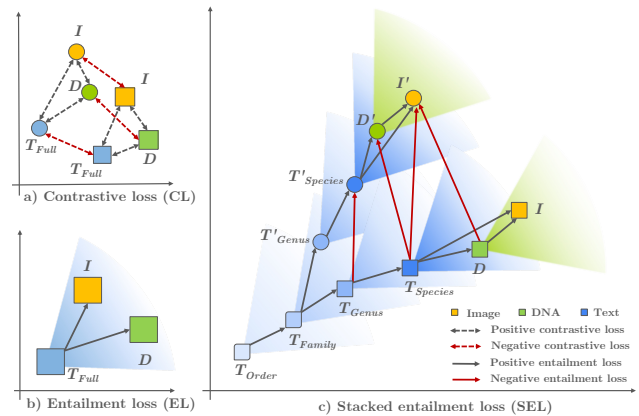

Figure 1. (a) **Contrastive loss**: instance-level alignment between modalities. (b) **Entailment loss**: enforces hierarchy within the text modality using entailment cones. (c) **Stacked entailment loss**: combines EL and cross-modal constraints by aligning image and DNA embeddings to multiple levels of the text hierarchy.

their embeddings into a shared hyperbolic space to promote taxonomic alignment across modalities.

Our experimental results show that our hyperbolic multimodal learning framework achieves strong performance in taxonomic classification and retrieval, especially at higher taxonomic ranks. The approach consistently matches or outperforms Euclidean baselines and better preserves the hierarchical relationships among modalities. However, all methods—including ours—face challenges in fine-grained species classification, particularly for previously unseen taxa. These results highlight both the potential of hyperbolic learning for hierarchical biological data, and the ongoing difficulty of open-world classification for biodiversity.

2. Related Works

Euclidean Multimodal Learning is the norm for recent advances in the multimodal contrastive learning domain, both in general vision-language frameworks such as CLIP [15] and SigLIP [19] and domain-specific ones, including those for biodiversity applications [5, 6, 17]. These biodiversity models embed images, textual data, and optionally DNA

ICCV
#9

ICCV
#9

ICCV 2025 Submission #9. CONFIDENTIAL REVIEW COPY. DO NOT DISTRIBUTE.

barcodes into a shared Euclidean embedding space using modality-specific encoders and contrastive learning objectives. CLIBD [5] in particular demonstrates zero-shot classification on BIOSCAN-1M [4], achieving superior accuracy than unimodal baselines.

Hyperbolic Representation Learning is an approach that utilizes hyperbolic geometry to encode features into a hierarchical representation space [11]. Unlike Euclidean space, hyperbolic spaces grow exponentially, matching the way the number of nodes in a hierarchy can grow exponentially with the depth. Nickel and Kiela [12] showed that taxonomic relationships in language can be effectively captured using hyperbolic embeddings. Recently, hyperbolic visual representation learning has been applied to vision tasks such as image retrieval [8] and image segmentation [3]. While the majority of these works use hyperbolic geometry only at the last layer, recent advances have been made towards developing *fully* hyperbolic models, *e.g.*, Poincaré ResNet [18].

Hyperbolic Multimodal Learning combines multimodal learning and the use of hyperbolic geometry to co-align embeddings from different modalities in a hierarchical representation space. Early examples include the work of Liu et al. [10], who demonstrated how to align images and text embeddings in a Poincaré hyperbolic space. MERU [2] uses contrastive learning to align images and text in Lorentzian space. It has served as the foundation for several recent models such as HyCoCLIP [13], which incorporates compositional constraints to strengthen fine-grained alignment between parts and wholes in visual concepts, as well as machine unlearning approaches [14, 20]. These works demonstrate hyperbolic geometry can enhance the structural consistency and interpretability of multimodal models, particularly in settings with implicit or weakly defined hierarchies.

Our method differs from prior work in three key ways. First, rather than focusing on vision-language or object-level inputs, we incorporate biologically grounded modalities—DNA barcodes and taxonomic labels—that are more salient for species-level classification. Second, we leverage *explicit* taxonomic hierarchies to guide representation learning rather than relying on *implicit* hierarchical signals such as caption specificity or object part composition. Third, our stacked entailment loss enforces consistency across hierarchical ranks.

3. Approach

We propose a multimodal representation learning framework that unifies specimen DNA barcodes, images, and taxonomic labels into a shared *hyperbolic* embedding space. By leveraging hyperbolic geometry, we aim to preserve hierarchical taxonomic relationships, improving classification accuracy and representation quality across the hierarchy.

Our framework employs three specialized encoders to process each of the data modalities: an image encoder extracts visual features, a DNA encoder encodes genetic sequences, and a text encoder captures semantic information from taxonomic labels of varying depth. These encoders independently map their inputs into a common embedding space, in which contrastive learning aligns multimodal representations for downstream tasks. We expand on CLIBD [5] by lifting the embeddings into hyperbolic space, and evaluate on the BIOSCAN-1M dataset [4].

3.1. Input and Output Specification

During training, the model receives triplets of aligned data: an image of a biological specimen, a DNA barcode sequence, and structured taxonomic labels (e.g., "Order: Diptera; Family: Syrphidae; Genus: Episyrphus; Species: Episyrphus balteatus"). Each modality is passed through a modality-specific encoder and projected into a shared hyperbolic embedding space. The training objective aligns these representations while preserving taxonomic hierarchy.

At inference time, the model supports both uni- and cross-modal retrieval, allowing it to taxonomically classify specimens using any available combination of images, DNA barcodes and taxonomic labels. This enables robust downstream use in biodiversity discovery and classification, even with missing or noisy modalities.

3.2. Encoders

Our framework adopts a modular design with modality-specific encoders that map images, DNA barcodes, and taxonomic text into a shared hyperbolic embedding space. We adapt the experimental setup from Gong et al. [5], using pretrained ViT-B/16, BERT-Small, and BarcodeBERT encoders for image, text, and DNA barcode modalities.

Each encoder produces Euclidean embeddings, which are then projected into a Lorentzian hyperbolic space with curvature c, via an exponential mapping centred at the origin. We refer the reader to Desai et al. [2] for details. The shared space enables contrastive alignment across modalities while preserving the hierarchical taxonomic structure.

3.3. Stacked Entailment Loss

To better leverage the inherent structure of the biological taxonomy, we propose a hierarchical learning objective termed *stacked entailment loss* (SEL). This mechanism is designed to explicitly enforce geometric relationships between taxonomic ranks—order, family, genus, and species—within hyperbolic space (see Figure 1). The design is inspired by compositional entailment mechanisms introduced in prior work [13], but adapted to reflect the nested and non-overlapping nature of biological hierarchies.

The core idea is to constrain the embeddings of lower-level taxa (e.g., genus) to lie within an entailment cone of their parent nodes (e.g., family). This entailment constraint is applied between each consecutive pair of levels in

ICCV
#9

ICCV 2025 Submission #9. CONFIDENTIAL REVIEW COPY. DO NOT DISTRIBUTE.

ICCV
#9

the hierarchy to ensure each child node is within the space "above" its parent, using a margin-based loss. Additionally, we introduce a *negative entailment* loss term which ensures each child node is *not* within the space "above" nodes from the preceding layer that are *not* its parent.

In general, given a batch $\mathcal{B} = \{(x_i, y_i, c_i)\}_{i=1}^{B}$, where x_i and y_i are embeddings and c_i the class, we define positive pairs $\mathcal{P} = \{(i,j) : c_i = c_j\}$ and negative pairs $\mathcal{N} = \{(i,j) : c_i \neq c_j\}$. The corresponding entailment losses are:

$$L_{\text{ent}}^+ = \frac{1}{|\mathcal{P}|} \sum_{(i,j) \in \mathcal{P}} \max\left(0, \, \text{ext}(x_i, y_j) - \text{aper}(x_i)\right) \quad (1)$$

$$L_{\text{ent}}^- = \frac{1}{|\mathcal{N}|} \sum_{(i,j) \in \mathcal{N}} \max\left(0, \, \text{aper}(x_i) - \text{ext}(x_i, y_j) + m\right) \quad (2)$$

where $\text{ext}(x, y)$ denotes the exterior angle between x and y in hyperbolic space, $\text{aper}(x)$ is the cone aperture of x, and m is the margin for negative pairs. The positive and negative entailment loss are then combined into: $L_{\text{ent}} = {}^1/_2 \left(L_{\text{ent}}^+ + L_{\text{ent}}^-\right)$. Unlike flat contrastive objectives, which treat all positive pairs equally, the stacked entailment loss introduces a directional notion of containment, ensuring that more specific taxa (fine-grained nodes) are properly nested under their broader ancestors in the hyperbolic hierarchy. The overall stacked-entailment loss consists of two parts: $L_{\text{SEL}} = L_{\text{SEL-intra}} + L_{\text{SEL-inter}}$. The first component, **intra-modal entailment loss**, enforces hierarchy among taxonomic labels. Let the taxonomy have R levels (e.g., order, family, genus, species), indexed $r = 1, 2, \ldots, R$ from root to leaf. T_r is the embedding at rank r, and $\mathbb{1}_r$ an indicator function for the availability of the label at rank r. Then we construct the intra-modal stacked entailment loss,

$$L_{\text{SEL-intra}} = \frac{1}{\sum_{r=2}^{R} \mathbb{1}_r} \sum_{r=2}^{R} \mathbb{1}_r \times L_{\text{ent}}(T_r, T_{r-1}). \quad (3)$$

Secondly, we introduce an **inter-modal entailment loss** that bridges the taxa labels with other modalities:

$$L_{\text{SEL-inter}} = \frac{1}{2} \left(L_{\text{ent}}(I, T_{R'}) + L_{\text{ent}}(D, T_{R'}) + L_{\text{ent}}(I, D) \right) \quad (4)$$

where I and D are the embeddings of images and DNA barcodes respectively, and $T_{R'}$ refers to the deepest available taxonomy label (*i.e.* T_{Species} if species is known, T_{Genus} if species isn't known but genus is, *etc.*). This term ensures that modality-specific inputs are not only aligned with the correct label, but also geometrically nested within the same hierarchical space. Since there can be multiple specimens with the same DNA barcode, and repeated photography of even the same specimen will not yield the same image, we consider the barcode to be more abstract than the image and also include an entailment loss term from barcode to image in the inter-modality objective.

In summary, our stacked entailment loss unifies taxonomic ordering and modality alignment, and embeds hierarchical structure into model training. This enables better generalization, especially with incomplete labels or unseen species. By explicitly modelling the hierarchical containment of taxonomic levels, our approach enables independent retrieval and prediction at any rank (e.g., order, family, genus, or species), facilitating multi-level querying and evaluation directly within the learned representation. This stands in contrast to CLIBD, which requires simultaneous prediction of all taxonomic levels during inference. We also extend the stacked entailment loss with two variants.

- **Image-DNA contrastive loss:** By adding a contrastive loss term based on the negative Lorentz distance between image and DNA embeddings, we encourage stronger cross-modal alignment and can improve the accuracy of image-to-DNA retrieval.
- **Full-text supervision:** We introduce an extra language input by concatenating taxonomic labels from all four ranks (order, family, genus, species), as is used in CLIBD. The full text is also used for contrastive alignment to the image and DNA embeddings.

4. Experiments and Results

We use the Euclidean-space CLIBD model [5] as a baseline, and adapt the CLIBD training pipeline to use hyperbolic-space based on the MERU framework [2]. We experimented with different combinations of loss functions, including entailment loss, stacked entailment loss, and contrastive loss. Experiments were conducted on four NVIDIA A100 GPUs (80GB VRAM each). We use a batch size of 2000 (4 × 500), except for experiments using stacked entailment, which could only fit a batch size of 1520 (4 × 380). All models were trained for 50 epochs with the Adam optimizer [9]. The learning rate was scheduled using a one-cycle policy [16], ranging from 1e-6 to 5e-5. We also used a trainable temperature for the contrastive loss, initialized to 0.07.

4.1. Metrics and Datasets

We conduct experiments on the BIOSCAN-1M dataset [4], which provides high-quality images with paired DNA barcodes and taxonomic labels for over 1 million insect specimens. For simplicity, we train all models on CLIBD's `train_seen` split of BIOSCAN-1M (36 k samples), which ensures all samples have complete species-level labels. We leave expanding the experiments to the full BIOSCAN-1M training dataset to future work.

Similar to CLIBD we evaluate classification performance across taxonomic ranks and for both seen and unseen classes, using class-averaged (macro) top-1 accuracy.

4.2. Results

We compare our hyperbolic SEL strategy against baselines on the BIOSCAN-1M dataset across three retrieval tasks (DNA-to-DNA, Image-to-Image, and Image-to-DNA) evaluated at four taxonomic levels (order, family, genus, and

ICCV
#9

ICCV
#9

ICCV 2025 Submission #9. CONFIDENTIAL REVIEW COPY. DO NOT DISTRIBUTE.

Table 1. Macro top-1 accuracy (%) comparison of different training objectives across taxonomic levels on BIOSCAN-1M. CL: contrastive loss. EL: entailment loss; SEL: stacked entailment loss; We evaluate uni- and multi-modal retrieval tasks including DNA-to-DNA, Image-to-Image, and Image-to-DNA. Accuracy is reported on both seen and unseen taxa, along with their harmonic mean (H.M.). Each method is further characterized by the configuration of entailment loss used (EL config.), whether full taxonomic text embedding is included utilized during training (Full Text), and the choice of embedding space (Euclidean: $\mathbb{R}^n$, or Lorentzian-hyperbolic: $\mathbb{H}^n_L$). All models are trained on the `train_seen` split of CLIBD and evaluated on the `test` split. **Best** results are shown in bold; <u>second-best</u> are underlined.

Rank	Method	EL config.	Full Text	Space	DNA-to-DNA			Image-to-Image			Image-to-DNA		
					Seen	Unseen	H.M.	Seen	Unseen	H.M.	Seen	Unseen	H.M.
Order	CLIBD	–	✓	$\mathbb{R}^n$	**89.1**	87.8	88.4	**99.5**	**66.4**	**79.6**	**98.7**	**49.5**	**65.9**
	CL	–	✓	$\mathbb{H}^n_L$	**89.1**	85.6	87.3	98.5	61.2	75.5	89.1	47.8	62.2
	EL+CL	Pos.	✓	$\mathbb{H}^n_L$	88.6	86.5	87.5	98.6	56.9	72.1	77.8	48.4	59.7
	SEL	Pos.+Neg.	✗	$\mathbb{H}^n_L$	88.4	**90.8**	**89.6**	79.3	62.3	69.8	**98.7**	<u>48.9</u>	<u>65.4</u>
	SEL+CL	Pos.+Neg.	✗	$\mathbb{H}^n_L$	88.7	86.3	87.5	<u>99.4</u>	<u>65.9</u>	<u>79.3</u>	78.6	48.2	59.7
	SEL+CL	Pos.+Neg.	✓	$\mathbb{H}^n_L$	88.9	<u>88.2</u>	88.5	99.0	60.9	75.4	78.6	<u>48.9</u>	60.3
Family	CLIBD	–	✓	$\mathbb{R}^n$	<u>90.8</u>	75.8	82.6	**89.2**	**52.2**	**65.9**	**83.6**	**19.3**	**31.4**
	CL	–	✓	$\mathbb{H}^n_L$	90.3	76.6	<u>82.9</u>	<u>83.9</u>	<u>48.5</u>	<u>61.4</u>	<u>79.6</u>	<u>18.8</u>	<u>30.4</u>
	EL+CL	Pos.	✓	$\mathbb{H}^n_L$	89.3	74.9	81.4	81.9	37.6	51.5	76.7	16.8	27.6
	SEL	Pos.+Neg.	✗	$\mathbb{H}^n_L$	86.8	**78.8**	82.6	79.0	41.8	54.7	78.9	18.4	29.9
	SEL+CL	Pos.+Neg.	✗	$\mathbb{H}^n_L$	89.0	76.9	82.5	79.6	46.6	58.8	78.7	17.3	28.4
	SEL+CL	Pos.+Neg.	✓	$\mathbb{H}^n_L$	**91.2**	<u>77.0</u>	**83.6**	82.4	41.5	55.2	78.1	17.4	28.4
Genus	CLIBD	–	✓	$\mathbb{R}^n$	85.2	64.3	73.3	**71.3**	**35.0**	**47.0**	**70.8**	**7.1**	**12.9**
	CL	–	✓	$\mathbb{H}^n_L$	**86.4**	64.9	<u>74.1</u>	<u>65.6</u>	32.4	43.4	66.9	6.5	11.8
	EL+CL	Pos.	✓	$\mathbb{H}^n_L$	84.7	63.1	72.3	63.0	22.8	33.5	64.2	<u>6.6</u>	11.9
	SEL	Pos.+Neg.	✗	$\mathbb{H}^n_L$	82.7	<u>65.9</u>	73.4	62.1	29.2	39.7	63.1	<u>6.6</u>	<u>12.0</u>
	SEL+CL	Pos.+Neg.	✗	$\mathbb{H}^n_L$	83.6	**66.9**	**74.3**	63.3	<u>33.1</u>	<u>43.5</u>	<u>67.6</u>	6.4	11.7
	SEL+CL	Pos.+Neg.	✓	$\mathbb{H}^n_L$	<u>85.8</u>	64.8	73.9	64.8	27.5	38.6	64.8	6.2	11.4
Species	CLIBD	–	✓	$\mathbb{R}^n$	81.8	60.6	69.7	**55.1**	**24.3**	**33.7**	**55.8**	0.7	1.4
	CL	–	✓	$\mathbb{H}^n_L$	**84.4**	61.8	**71.4**	<u>48.2</u>	22.6	<u>30.8</u>	53.7	0.9	1.7
	EL+CL	Pos.	✓	$\mathbb{H}^n_L$	82.5	60.1	69.6	45.4	14.3	21.8	50.5	0.9	1.8
	SEL	Pos.+Neg.	✗	$\mathbb{H}^n_L$	79.5	<u>62.3</u>	69.9	45.5	20.0	27.8	52.0	**1.1**	**2.1**
	SEL+CL	Pos.+Neg.	✗	$\mathbb{H}^n_L$	80.5	**63.2**	<u>70.8</u>	46.8	<u>22.8</u>	30.7	<u>54.2</u>	0.7	1.4
	SEL+CL	Pos.+Neg.	✓	$\mathbb{H}^n_L$	<u>82.6</u>	62.0	<u>70.8</u>	47.8	19.0	27.2	51.4	<u>1.0</u>	**2.1**

species). We investigate how well training with contrastive loss (CL) in the hyperbolic space performs compared with training in Euclidean space (CLIBD [5]). We then compare different ways of training in hyperbolic space, comparing a strategy similar to MERU [2] with entailment loss and contrastive losses (EL + CL) to different variants of SEL. Table 1 reports macro Top-1 accuracy for seen and unseen taxa, as well as their harmonic mean.

Across all retrieval tasks, models achieve high accuracy at the coarsest levels, but this falls off substantially as ranks become more fine-grained, especially for image-based retrieval. We note that hyperbolic models consistently achieve results that are highly comparable to the Euclidean CLIBD baseline across all ranks and retrieval settings. SEL methods consistently perform best at unseen DNA retrieval, whereas the Euclidean model performs best at image retrieval. Comparing EL+CL to SEL+CL with full text, we find that SEL+CL always dominates the former, demonstrating the utility of the stacked entailment over single-layer entailment. Additionally, comparing SEL+CL without full text to with it, we find this change consistently improves unimodal seen taxa retrieval, but decreases unseen taxa performance and cross-modal performance.

5. Discussion

Our experiments demonstrate hyperbolic learning can effectively capture hierarchical structure in biological data and provides performance competitive with established Euclidean methods. However, neither framework fully overcomes the persistent challenge of fine-grained, open-world species identification.

Improving classification at fine-grained taxonomic ranks and for novel, unseen taxa remains a key direction for future work. Potential strategies include addressing class imbalance, enhancing data augmentation, or leveraging more advanced hierarchical or uncertainty-aware methods.

ICCV
#9

ICCV
#9

ICCV 2025 Submission #9. CONFIDENTIAL REVIEW COPY. DO NOT DISTRIBUTE.

References

[1] Thomas WA Braukmann, Natalia V Ivanova, Sean WJ Prosser, Vasco Elbrecht, Dirk Steinke, Sujeevan Ratnasingham, Jeremy R de Waard, Jayme E Sones, Evgeny V Zakharov, and Paul DN Hebert. Metabarcoding a diverse arthropod mock community. *Molecular ecology resources*, 19(3):711–727, 2019, doi:10.1111/1755-0998.13008. 1

[2] Karan Desai, Maximilian Nickel, Tanmay Rajpurohit, Justin Johnson, and Shanmukha Ramakrishna Vedantam. Hyperbolic image-text representations. In *International Conference on Machine Learning*, pages 7694–7731. PMLR, 2023. 2, 3, 4

[3] 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*, pages 4453–4462, 2022. 2

[4] Zahra Gharaee, ZeMing Gong, Nicholas Pellegrino, Iuliia Zarubiieva, Joakim Bruslund Haurum, Scott C. Lowe, Jaclyn T. A. McKeown, Chris C. Y. Ho, Joschka McLeod, Yi-Yun C. Wei, Jireh Agda, Sujeevan Ratnasingham, Dirk Steinke, Angel X. Chang, Graham W. Taylor, and Paul Fieguth. A step towards worldwide biodiversity assessment: The BIOSCAN-1M insect dataset. In *Advances in Neural Information Processing Systems*, pages 43593–43619. Curran Associates, Inc., 2023. 2, 3

[5] ZeMing Gong, Austin T. Wang, Xiaoliang Huo, Joakim Bruslund Haurum, Scott C. Lowe, Graham W. Taylor, and Angel X. Chang. CLIBD: Bridging vision and genomics for biodiversity monitoring at scale. In *The Thirteenth International Conference on Learning Representations*, 2025, doi:10.48550/arxiv.2405.17537. 1, 2, 3, 4

[6] Jianyang Gu, Samuel Stevens, Elizabeth G Campolongo, Matthew J Thompson, Net Zhang, Jiaman Wu, Andrei Kopanev, Zheda Mai, Alexander E. White, James Balhoff, Wasila Dahdul, Daniel Rubenstein, Hilmar Lapp, Tanya Berger-Wolf, Wei-Lun Chao, and Yu Su. BioCLIP 2: Emergent properties from scaling hierarchical contrastive learning. *arXiv preprint arXiv:2505.23883*, 2025, doi:10.48550/arxiv.2505.23883. 1

[7] Paul DN Hebert, Alina Cywinska, Shelley L Ball, and Jeremy R DeWaard. Biological identifications through DNA barcodes. *Proceedings of the Royal Society of London. Series B: Biological Sciences*, 270(1512):313–321, 2003, doi:10.1098/rspb.2002.2218. 1

[8] Valentin Khrulkov, Leyla Mirvakhabova, Evgeniya Ustinova, Ivan Oseledets, and Victor Lempitsky. Hyperbolic image embeddings. In *The IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020. 2

[9] Diederik Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *International Conference on Learning Representations (ICLR)*, 2015, doi:10.48550/arXiv.1412.6980. 3

[10] Shaoteng Liu, Jingjing Chen, Liangming Pan, Chong-Wah Ngo, Tat-Seng Chua, and Yu-Gang Jiang. Hyperbolic visual embedding learning for zero-shot recognition. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020. 2

[11] Pascal Mettes, Mina Ghadimi Atigh, Martin Keller-Ressel, Jeffrey Gu, and Serena Yeung. Hyperbolic deep learning in computer vision: A survey. *International Journal of Computer Vision*, 132(9):3484–3508, 2024, doi:10.1007/s11263-024-02043-5. 2

[12] Maximillian Nickel and Douwe Kiela. Poincaré embeddings for learning hierarchical representations. *Advances in neural information processing systems*, 30, 2017. 2

[13] Avik Pal, Max van Spengler, Guido Maria D'Amely di Melendugno, Alessandro Flaborea, Fabio Galasso, and Pascal Mettes. Compositional entailment learning for hyperbolic vision-language models. *arXiv preprint arXiv:2410.06912*, 2024, doi:10.48550/arxiv.2410.06912. 2

[14] Tobia Poppi, Tejaswi Kasarla, Pascal Mettes, Lorenzo Baraldi, and Rita Cucchiara. Hyperbolic safety-aware vision-language models. In *Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR)*, pages 4222–4232, 2025. 2

[15] 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*, pages 8748–8763. PMLR, 2021. 1

[16] Leslie N. Smith. A disciplined approach to neural network hyper-parameters: Part 1 - learning rate, batch size, momentum, and weight decay. *arXiv preprint arXiv:1803.09820*, 2018, doi:10.48550/arXiv.1803.09820. 3

[17] 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)*, pages 19412–19424, 2024. 1

[18] Max van Spengler, Erwin Berkhout, and Pascal Mettes. Poincare ResNet. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 5419–5428, 2023. 2

[19] Xiaohua Zhai, Basil Mustafa, Alexander Kolesnikov, and Lucas Beyer. Sigmoid loss for language image pre-training. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 11975–11986, 2023. 1

[20] Àlex Pujol Vidal, Sergio Escalera, Kamal Nasrollahi, and Thomas B. Moeslund. Machine unlearning in hyperbolic vs. Euclidean multimodal contrastive learning: Adapting alignment calibration to MERU. *arXiv preprint arXiv:2503.15166*, 2025, doi:10.48550/arxiv.2503.15166. 2