Learning Visual Hierarchies in Hyperbolic Space for Image Retrieval

Ziwei Wang[12*] Sameera Ramasinghe[1*] Chenchen Xu[1]
Julien Monteil[1] Loris Bazzani[1] Thalaiyasingam Ajanthan[1*]

[1]Amazon [2]Australian National University

Abstract

Structuring latent representations in a hierarchical manner enables models to learn patterns at multiple levels of abstraction. However, most prevalent image understanding models focus on visual similarity, and learning visual hierarchies is relatively unexplored. In this work, for the first time, we introduce a learning paradigm that can encode user-defined multi-level complex visual hierarchies in hyperbolic space without requiring explicit hierarchical labels. As a concrete example, first, we define a part-based image hierarchy using object-level annotations within and across images. Then, we introduce an approach to enforce the hierarchy using contrastive loss with pairwise entailment metrics. Finally, we discuss new evaluation metrics to effectively measure hierarchical image retrieval. Encoding these complex relationships ensures that the learned representations capture semantic and structural information that transcends mere visual similarity. Experiments in part-based image retrieval show significant improvements in hierarchical retrieval tasks, demonstrating the capability of our model in capturing visual hierarchies.

1. Introduction

Humans organize knowledge of the world into hierarchies [26] for efficient knowledge management. Developing models that encode such hierarchies is crucial for creating systems with holistic world understanding aligned with human perception. While this topic spans various modalities, we focus on encoding hierarchies in the visual domain. For many large image datasets, objects can be organized according to latent hierarchies [26], as evidenced by power law distributions [34]. However, most prevalent image understanding models [4, 14–16, 36] focus on preserving visual similarity, and consequently, learning hierarchies in the visual domain remains relatively unexplored. These limitations hinder models' ability to generalize to tasks requiring hierarchical reasoning such as understanding scenes, object parts, and their interactions at multiple levels of abstraction. We illustrate in Fig. 1 an example that shows the complexity of a visual hier-

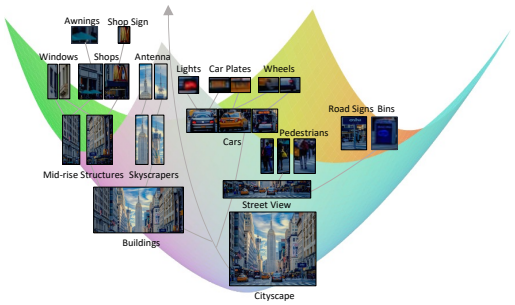


Figure 1. **An illustration of part-based image hierarchy organized in hyperbolic space.** At the highest level, we see the urban environment, composed of buildings, streets, and sky. Zooming in, we find the building category, which further divides into skyscrapers, mid-rise structures, and more. Each of them has its own visual elements, which in turn can be decomposed into sub-elements.

[*] work done while at Amazon

archy where elements share similarities both within and across categories.

Recent works have demonstrated the utility of hyperbolic representation spaces for capturing hierarchical relationships in an unsupervised setting [6, 30, 32]. This *emergent* latent structure is appealing; nevertheless, meaningful hierarchies are often task and data dependent, and aligning model behavior with such human-defined hierarchies is essential for many applications. To this end, we introduce a learning paradigm that can encode multi-level hierarchies as entailment pairs in hyperbolic space. As a concrete example, we first define a general part-based image hierarchy using object and part level classification annotations within and across images. Then, we introduce a model capable of structuring the latent space to preserve the defined hierarchy using only image/object level information. To our knowledge, we are the first to encode multi-level complex visual hierarchies without relying on explicit hierarchical labels or additional modalities. Finally, we introduce an evaluation metric to effectively measure hierarchical image retrieval.

Note that, hierarchy is an asymmetric relationship and has a high branching factor (see Fig. 1). To this end, we adopt the hyperbolic geometry as it provides a continuous approximation of such tree-like structures [26, 35]. To enforce the hierarchy, we break it into pairwise entailment relationships between images, objects, and parts, at multiple levels within an image as well as across images at category level. For pairwise entailment (asymmetric), we adapt the recently proposed angle-based asymmetric distance metric [32] within the contrastive learning paradigm, and extend it to handle cases with multiple positive relationships. In contrast to symmetric distances such as the inner-product used in [11, 20, 27], this angle-based distance offers an additional degree of freedom along the radial axis to form emergent structures in the latent space.

For experimentation, we build a dataset of visual hierarchies using the bounding box annotations of OpenImages [23]. Our dataset includes entailment relationships between scenes, objects, and parts, within a single image as well as across images at the category level. Similarly, for hierarchical retrieval evaluation, we use the full training set to create ground truth hierarchy trees per scene/object. Additionally, we design a metric for evaluating hierarchical retrieval based on the optimal transport distance between the label distribution of the retrieval set and ground truth label distribution within the hierarchy tree. Combined with Recall@k metrics, this demonstrates that our method captures semantic and structural information, transcending mere visual similarity. Furthermore, to the best of our knowledge, our model is the first to generalize to out-of-domain image hierarchies, achieving strong performance on unseen and out-of-domain datasets.

Our contributions can be summarized as follows:

- To our knowledge, for the first time, we introduce a new learning paradigm to effectively encode user-defined multi-level complex visual hierarchies in hyperbolic space that does not require explicit hierarchical labels.
- We adapt a contrastive loss using hyperbolic angle-based distance metric to enforce pairwise entailment relationships, and empirically demonstrate that pairwise entailment is sufficient to learn complex visual hierarchies.
- We introduce an optimal transport based evaluation metric to measure hierarchical image retrieval performance.
- We demonstrate superior generalization capabilities of our model beyond the user-defined hierarchies via out-of-domain unseen data evaluation and ablation experiments.

2. Preliminaries

We briefly review essential concepts in hyperbolic geometry here. We refer the reader to [33] for a comprehensive treatment. Hyperbolic spaces are Riemannian manifolds with constant negative curvature and are fundamentally different from Euclidean or spherical space which has zero or constant positive curvature, respectively. The negative curvature enables properties such as divergence of parallel lines and exponential volume growth with radius [2]. This volume growth property makes hyperbolic space an ideal candidate for embedding hierarchical and graph structured data [26], and has found many machine learning applications.

Lorentz Model. The Lorentz model is a way to represent a hyperbolic space. It embeds the d-dimensional hyperbolic space $\mathbb{H}^d$ with curvature c within an $(d+1)$-dimensional Minkowski space $\mathbb{R}^{d,1}$, a pseudo-Euclidean space with one negative dimension, as follows:

$$\mathbb{H}^d = \left\{ \mathbf{x} \in \mathbb{R}^{d,1} \mid \langle \mathbf{x}, \mathbf{x} \rangle_{\mathbb{H}} = -1/c, x_0 > 0 \right\} , \tag{1}$$

where the Lorentzian inner product is defined as,

$$\langle \mathbf{x}, \mathbf{y} \rangle_{\mathbb{H}} = -x_0 y_0 + \sum_{i=1}^{d} x_i y_i . \tag{2}$$

Here, the 0-th dimension of the vector is treated as the time component and the rest as the space component. From the definition of $\mathbb{H}^d$, the time component can be written using the space component as follows:

$$x_{\text{time}} = x_0 = \sqrt{1/c + \|\mathbf{x}_{\text{space}}\|^2} , \tag{3}$$

where $\|\cdot\|$ is the Euclidean norm and $\mathbf{x}_{\text{space}} = \mathbf{x}_{1:d}$.

Tangent Spaces. The tangent space at a point $\mathbf{x} \in \mathbb{H}^d$ in hyperbolic space, is a Euclidean space that locally approximates hyperbolic space around $\mathbf{x}$. Exponential and logarithmic maps are used to project a point from a tangent space to hyperbolic space and vice versa.

3. Enforcing User-Defined Hierarchies

Our aim is to define a hierarchy in images and enforce it in the latent space using hyperbolic geometry. We first define a part-based hierarchy in images, then discuss our approach to enforce it, and finally introduce our hierarchical retrieval metric.

3.1. Part-Based Image Hierarchy

Visual hierarchies can be established in different ways, depending on the application. In this work, we are interested in a hierarchy that encapsulates the semantic relationship among objects in a scene. For this, *scene-object-part* hierarchy is appealing as it is useful for applications such as fine-grained object retrieval, object localization, and general scene understanding. This hierarchy has also been shown to emerge in hyperbolic image embeddings [20, 32].

Given an image dataset with bounding box and object class annotations, we define a part-based image hierarchy where the full scene images – typically containing multiple objects – represent the highest level in the hierarchy, while individual objects constitute progressively lower levels, *entailed* by the full image. In this, larger bounding boxes that significantly envelope smaller ones are considered to entail those smaller ones, establishing a nested hierarchy. In this way, from the full scene to the smallest bounding box, a hierarchy can be defined by recursively applying the entailment rule: *if B contained in A, then A entails B, denoted as $A \rightarrow B$*. An illustrative example is shown in Fig. 2, where an example hierarchy could be `road scene` $\rightarrow$ `cyclist` $\rightarrow$ `bicycle` $\rightarrow$ `wheels`.

Pairwise Entailment. Our entailment rule above naturally facilitates a pairwise relationship. Let $I \in \mathcal{I}$ be an image, either the full scene image or a cropped bounding box, and let $\mathcal{B}$ denote the set of all bounding boxes in the dataset. Suppose $\mathcal{B}_I$ be the set of bounding boxes contained in the image I, *i.e.*, $\mathcal{B}_I = \{b \in \mathcal{B} \mid b \subset I\}$[2]. Note, each bounding box has an associated object label denoted by $b_l \in \mathcal{L}$.

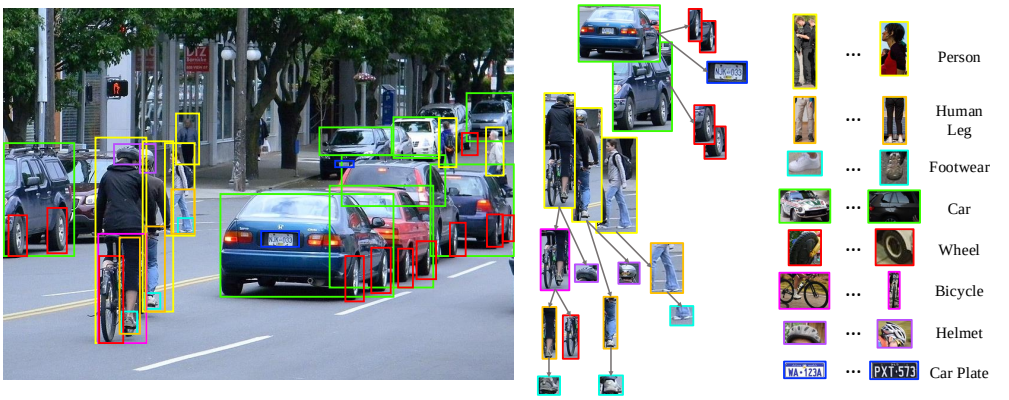
We define the following pairwise entailment relationship:

$$\text{if} \;\; b \in \mathcal{B}_I, \;\; \text{then} \;\; I \rightarrow b . \tag{4}$$

By applying this recursively, a tree-like hierarchy can be formed as shown in Fig. 2b. Note that our model can encode any "user-defined hierarchy" represented as entailment pairs in Eq. (4). Part-based image hierarchy is one use case.

Furthermore, if I is a full scene image, we define an additional entailment relationship across images at the object level. Specifically, for each bounding box b in the image I, we sample K bounding boxes from other images with the same label b_l and enforce entailment with the image I. Formally, let $a \in \mathcal{L}$, and let $\mathcal{B}^a_{I,K} \sim \{b \in B \mid b \not\subset I, b_l = a\}$ be the set of K bounding boxes of label a on images other

[2]We use the $\subset$ notation to denote *contained in* relationship. In the case where I is a cropped bounding box, this relationship is defined to hold if the majority (*e.g.*, 80%) of the small bounding box b is contained in I.



(a). Parent image with bounding boxes (b). Example hierarchies (c). Cross-image samples for each class

Figure 2. **An illustrative example image hierarchies.** a) Image I with object-level bounding boxes. Each bounding box is entailed by I. b) Hierarchies created via bounding box-to-bounding box entailment within I (larger bounding boxes entail smaller ones). c) Cross-image hierarchy created by sampling N bounding boxes with corresponding object classes from other images, which are then entailed by I. Find details of cross-image sampling in Sec. 3.1.

than I. We enforce the entailment as follows:

$$\text{for all } \; b \in \mathcal{B}_I, \;\; \text{if } \; x \in \mathcal{B}_{I,K}^{b_l}, \;\; \text{then } \; I \rightarrow x \, . \quad (5)$$

This additional entailment across images reinforces the semantic link between scenes and similar objects across different images (see Fig. 2c).

We posit that these relationships help to structure a hierarchical understanding of images based on scene, object, and part relationships.

Hierarchy Tree. As noted above, the part-based hierarchy forms a tree structure, where an image or cropped bounding box can be traversed using pairwise entailment relationships. For evaluating hierarchical image retrieval, we construct this hierarchy tree per scene/object automatically using the full training set. However, the model is trained solely on pairwise entailment relationships and does not use the hierarchy tree.

3.2. Angle-Based Entailment Loss

We require an asymmetric distance function to enforce pairwise entailment relationships in hyperbolic space. To this end, we adapt the recently proposed hyperbolic-angle-based entailment loss [32], a smooth contrastive variant of the entailment cone loss [8]. The angle-only loss, without distance constraints on image pairs, provides flexibility to be distributed along the radial axis, allowing embeddings to align with the tree structure while preserving pairwise entailment.

Our loss is a bidirectional version of [32], as illustrated graphically in Fig. 3. In particular, given embeddings of an entailment pair, $\mathbf{x}, \mathbf{y} \in \mathbb{R}^d$ in the tangent space, where $\mathbf{x}$ entails $\mathbf{y}$, we maximize the angles β_1 and α_2. This enforces entailment in a bidirectional manner. The angles β_1 and α_2 can be computed using the exterior angle as follows:

$$\begin{aligned} \beta_1(\mathbf{x}, \mathbf{y}) &= \pi - \text{ext}(\mathbf{x}, \mathbf{y}) \, , \\ \alpha_2(\mathbf{y}, \mathbf{x}) &= \text{ext}(\mathbf{y}, \mathbf{x}) \, , \end{aligned} \quad (6)$$

where $\text{ext}(\mathbf{x}, \mathbf{y})$ is the exterior angle between $\mathbf{x}$ and $\mathbf{y}$ and takes the following form [6]:

$$\text{ext}(\mathbf{x}, \mathbf{y}) = \cos^{-1}\left(\frac{y_{\text{time}} + x_{\text{time}} \, c\langle \mathbf{x}, \mathbf{y} \rangle_{\mathbb{H}}}{\|\mathbf{x}_{\text{space}}\| \sqrt{\left(c\langle \mathbf{x}, \mathbf{y} \rangle_{\mathbb{H}}\right)^2 - 1}} \right) . \quad (7)$$

In contrast to [32], in our case an embedding can belong to multiple entailment pairs in a batch. This corresponds to a case of multiple positives in the contrastive loss. Thus, we employ the InfoNCE loss [28] to align all entailment pairs while pushing apart the rest of the negative pairs. Precisely, let $\mathcal{D} = \{(\mathbf{x}_i, \mathbf{y}_i)\}$ be a batch of entailment pairs, then

the InfoNCE loss for parent-to-child can be written as:

$$L^{p\to c}(\mathcal{D},\kappa) = \tag{8}$$
$$-\mathbb{E}_{\mathcal{D}}\left[\log \frac{\exp\left(\frac{\kappa(\mathbf{x}_i,\mathbf{y}_i)}{\tau}\right)}{\exp\left(\frac{\kappa(\mathbf{x}_i,\mathbf{y}_i)}{\tau}\right)+\sum_{\mathbf{y}^-\in\mathcal{N}_i}\exp\left(\frac{\kappa(\mathbf{x}_i,\mathbf{y}_i^-)}{\tau}\right)}\right],$$

where $\mathcal{N}_i$ denotes the set of samples that do not have an entailment relationship with the parent embedding $\mathbf{x}_i$. Here, $\kappa : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}$ is the similarity function, and τ is a learnable temperature parameter initialized to 0.07 following [6]. Now, our bidirectional entailment loss can be written as:

$$L_{\text{angle}}(\mathcal{D}) = L^{p\to c}(\mathcal{D},\beta_1) + L^{c\to p}(\mathcal{D},\alpha_2)\,. \tag{9}$$

Here, the similarity function κ is replaced with angles β_1 and α_2 so that the contrastive loss maximizes angles β_1 and α_2 for matching entailment pairs in the batch $\mathcal{D}$.

In our implementation, we use a shared image encoder for both parent and child embeddings. Following the reparametrization of [6], we encode the space component of the Lorentz model in the tangent space at origin and project it onto the hyperboloid using the exponential map, enabling contrastive entailment angle loss computation in hyperbolic space.

Loss in Euclidean Space. This entailment angle loss is general and can be effectively enforced in Euclidean space. In Euclidean space, the exterior angles are formulated as follows:

$$\text{ext}(\mathbf{x},\mathbf{y})_{\mathbb{E}} = \cos^{-1}\left(\frac{\|\mathbf{y}\|^2 - \|\mathbf{x}\|^2 - \|\mathbf{x}-\mathbf{y}\|^2}{2\|\mathbf{x}\|\cdot\|\mathbf{x}-\mathbf{y}\|}\right). \tag{10}$$

Now, the loss can be analogously derived.

3.3. Hierarchical Retrieval Evaluation

To evaluate retrieval performance on the hierarchy tree, we also introduce a metric that captures the label distribution in the dataset. This is important as different labels can have different numbers of instances and the standard metrics such as Recall@k is agnostic to it.

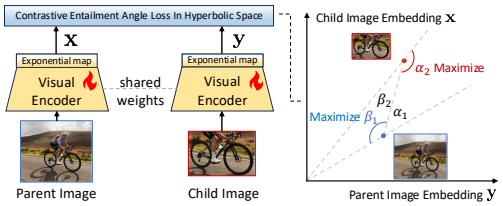


Figure 3. **Learning multi-level hierarchies via contrastive entailment angle loss.** Our model first encodes parent-to-child pairs into embeddings with exponential mapping, then maximizes β_1 and α_2 using our contrastive entailment angle loss in hyperbolic space.

To this end, consider the parent-to-child relationship, and let $\mathcal{H}_I$ denote the hierarchy tree originating from the query image I, containing m labels. Then, the labels in $\mathcal{H}_I$ are the ground truth labels for the hierarchical retrieval for image I. Now, let $\mathbf{h}_I \in \mathbf{R}^m$ be the precomputed label distribution of $\mathcal{H}_I$. Then, our optimal transport (1-D Wasserstein) distance between the retrieved label distribution $\mathbf{r}_I$ and $\mathbf{h}_I$ is:

$$\text{OT}(\mathbf{h}_I,\mathbf{r}_I) = \text{Wasserstein}(\bar{\mathbf{h}}_I,\bar{\mathbf{r}}_I)\,, \tag{11}$$

where $\bar{\cdot} \in \mathbf{R}^{m+1}$ is the label distribution with *other* class added[3]. Note that a smaller distance indicates better alignment.

4. Related Work

Hyperbolic geometry allows for exponential volume growth with respect to the radius [2], making it effective for embedding hierarchical structures. This advantage has led to significant research into leveraging hyperbolic representations for various data types, including molecular structures [41], 3D [3, 17, 39], images [11, 12, 20, 21, 30], text data [9, 18, 37, 37, 42], and vision-language data [1, 6, 22, 32].

Hyperbolic embeddings can be learned through standard deep learning layers [19] with hyperbolic projection [26] or using hyperbolic neural networks [9]. Many prior NLP, computer vision and knowledge graph studies learn hierarchies from partially order data [9, 24, 38], or minimizing geodesic distance or maximizing similarities [11, 26, 27, 40].

[3]For $\mathbf{h}_I$ other class has zero mass, and for $\mathbf{r}_I$ all labels not in $\mathcal{H}_I$ are combined to form the other class.

Ganea et al. [9] introduced an angle-based entailment cone loss which pushes child nodes into the cone emanating from the parent node embedding. This approach has been applied to both text [9] and image data [7] with label hierarchies. Recently, this hyperbolic entailment loss was adapted for contrastive learning to develop representations in vision-language models [1, 6, 29, 32]. However, such methods remain relatively unexplored in the image domain. We adapted the angle-based entailment loss from [32] to encode part-based image hierarchies. Many previous works are limited to learning hierarchies for single-class images using predefined labeled hierarchies, such as ImageNet [20, 25] or hand-labeled data [7]. In this work, we propose a learning method that fine-tunes pre-trained models on large-scale datasets for **general images** (with multiple classes per image) **without hierarchical labels**. The most relevant approach, HCL [11], models simple scene-to-object hierarchies, whereas we capture more complex, multi-level part-based hierarchies directly from image data, extending beyond visual similarity.

5. Experiments

Datasets. For training and evaluation, we construct *HierOpenImages*, a novel dataset containing pairwise part-based image hierarchies built from the OpenImages dataset [23]. We further evaluate generalization on out-of-domain datasets and hierarchies.

Models. We evaluate the performance of learning multi-level image hierarchies using two popular visual encoders 1) CLIP ViT (B/16) [31], pretrained on large-scale image-text pairs from the Internet, and 2) MoCo-v2 (ResNet-50) [5], pretrained on ImageNet. Note that both CLIP and MoCo-v2 models are fine-tuned and evaluated in an **image encoder only** setting. We use these pretrained image models as our baseline and compare our proposed angle-based hyperbolic method against its Euclidean counterpart. Each model is fine-tuned for a single epoch on *HierOpenImages* using the proposed contrastive angle-based entailment loss, and † denotes fine-tuning. The retrieval distance function (Dist. Func.) aligns with the scoring function used during training. For example, minimizing hyperbolic angles (Hyp Ang.) for CLIP-hyp† and Euclidean angles (Euc Ang.) for CLIP-euc† model.

We also compare the state-of-the-art hyperbolic hierarchical image-only model HCL [11], which is trained on scene-to-object relationships from the OpenImages dataset [23] in both hyperbolic and Euclidean space. Accordingly, we evaluate the retrieval performance of HCL [11] using both hyperbolic distance and cosine similarity. For completeness, we also fine-tuned HCL† [11] on the same *HierOpenImages* dataset and included it in our comparisons. Note that fine-tuning or evaluating with a text encoder is not possible on the current *HierOpenImages* dataset. Furthermore, previous image-only supervised methods, trained on predefined single-class image hierarchies [7, 20] are not directly comparable. These methods require all image classes to be predefined during training, which is not feasible for complex multi-object scenes in the OpenImages [23] and LVIS [13] datasets, where images contain multiple objects with diverse class labels.

5.1. Main Results

In same-class retrieval, we assess whether the retrieved image belongs to the *same class* as the query image. In hierarchical retrieval, we verify if the retrieved image exists within the *hierarchy tree* of the query image, specifically evaluating the quality of the learned hierarchical representations.

Same-Class Retrieval. Denoting the full image as *parent* and the bounding box as *child*, we evaluate retrieval tasks in both child-to-parent and parent-to-child directions. Table 1 shows the retrieval accuracy of top-k $= \{5, 10, 50, 100\}$. We notice a significant and consistent improvement with our proposed hyperbolic model, across all metrics and model variants. This highlights the relevance of our angle-based entailment loss and the advantages of learning hierarchical image embeddings in hyperbolic space. While for HCL [11], only a slight performance increase was observed after fine-tuning.

Hierarchical Retrieval via the Learned Latent Space Distribution. Table 2 evaluates the hierarchical structure of the latent space by retrieving a large number of child images from parent images.

Vision Encoder	Model	Metrics	Top-5	Top-10	Top-50	Top-100
Child-to-Parent						
CLIP ViT	CLIP	Cos Sim.	26.73	25.95	23.69	22.70
	CLIP-euc†	Euc Ang.	67.83	68.37	67.12	66.04
	CLIP-hyp†	Hyp Ang.	**73.37**	**72.59**	**69.84**	**68.62**
MoCo-v2	HCL	Cos Sim.	15.11	14.49	14.69	14.60
	HCL	Hyp Dist.	14.46	14.34	14.04	13.92
	HCL†	Hyp Dist.	14.54	14.43	14.16	14.01
	MoCo	Cos Sim.	17.23	16.86	16.33	16.07
	MoCo-euc†	Euc Ang.	54.51	54.16	**51.56**	**49.53**
	MoCo-hyp†	Hyp Ang.	**55.53**	**55.31**	51.53	49.51
Parent-to-Child						
CLIP ViT	CLIP	Cos Sim.	47.52	46.60	43.50	42.08
	CLIP-euc†	Euc Ang.	65.38	65.70	66.01	65.79
	CLIP-hyp†	Hyp Ang.	**66.02**	**66.63**	**66.50**	**65.91**
MoCo-v2	HCL	Cos Sim.	16.08	15.93	15.84	15.74
	HCL	Hyp Dist.	16.16	16.04	15.60	15.45
	HCL†	Hyp Dist.	17.07	16.57	15.97	15.69
	MoCo	Cos Sim.	18.49	18.37	17.73	17.45
	MoCo-euc†	Euc Ang.	47.11	46.86	46.95	47.22
	MoCo-hyp†	Hyp Ang.	**52.01**	**51.64**	**50.83**	**50.51**

Table 1. **Part-based same-class image retrieval evaluation.** For child-to-parent image retrieval, the retrieved parent must contain the object class of the query child. For parent-to-child image retrieval, the retrieved child must match a class within the parent. † denotes models fine-tuned on the *HierOpenImages* dataset. Our proposed method is shaded in purple.

Metrics	Model	Dist. Func.	Top-150k	Top-200k	Top-250k
CLIP ViT					
Recall ↑	CLIP	Cos Sim.	51.33	64.28	77.02
	CLIP-euc†	Euc Ang.	72.96	80.98	87.97
	CLIP-hyp†	Hyp Ang.	**74.06**	**82.34**	**88.79**
OT Distance ↓	CLIP	Cos Sim.	23.81	24.60	25.03
	CLIP-euc†	Euc Ang.	17.00	20.00	22.45
	CLIP-hyp†	Hyp Ang.	**16.36**	**19.27**	**22.21**
MoCo-v2					
Recall ↑	HCL	Cos Sim.	46.05	61.24	76.44
	HCL	Hyp Dist.	46.06	61.14	76.16
	HCL†	Hyp Dist.	45.81	60.77	75.77
	MoCo	Cos Sim.	48.38	63.63	77.89
	MoCo-euc†	Euc Ang.	55.81	**71.86**	**83.42**
	MoCo-hyp†	Hyp Ang.	**56.09**	71.61	83.13
OT Distance ↓	HCL	Cos Sim.	24.23	24.67	25.09
	HCL	Hyp Dist.	25.72	25.98	26.04
	HCL†	Hyp Dist.	26.12	26.32	26.28
	MoCo	Cos Sim.	24.05	24.44	24.93
	MoCo-euc†	Euc Ang.	13.72	16.83	20.17
	MoCo-hyp†	Hyp Ang.	**12.41**	**15.29**	**19.09**

Table 2. **Part-based hierarchical evaluation of parent-to-child image retrieval on *HierOpenImages*.** Results are evaluated using the ground truth hierarchy tree and the hierarchical distribution of the test set. Smaller OT distance indicates better distribution alignment.

We use recall to measure the percentage of ground truth images that are successfully retrieved. Moreover, we check the alignment between the retrieved distribution and the underlying hierarchical distribution of the full test set. Good distribution alignment is a desirable property for fine-grained retrieval as the *retrieved set should capture the hierarchies present in the data distribution*. We propose to measure distribution alignment using the optimal transport (Wasserstein distance), with a smaller distance indicating a closer match (see Sec. 3.3).

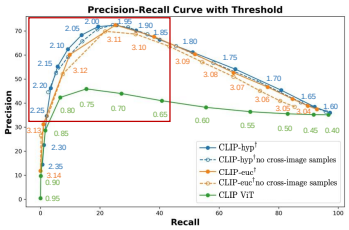


Figure 4. **Precision-Recall curves of CLIP ViT models on hierarchical retrieval.** Dotted lines show models trained only on hierarchical entailment data within the same images; solid lines represent models trained with additional cross-image scene-to-object samples. Angle or cosine similarity threshold values are marked by text.

As shown in Table 2, our hyperbolic model better captures the hierarchical distribution of the test set compared to the Euclidean model, achieving better OT distance in all cases, and in 4 out of 6 instances for the recall.

All fine-tuned models using our angle-based entailment loss show significant improvement over the baseline models. Notably, even after fine-tuning, HCL [11] shows a decline in hierarchical retrieval performance, indicating its difficulty in learning the complex visual hierarchies in the training data.

Effect of Enforcing Cross Image Entailment. To compare the emergent behaviors of the hyperbolic and Euclidean models, we trained the CLIP ViT model solely on hierarchical part-based entailment data within individual images (entailment pairs with high visual similarity), omitting any cross-image image-to-bounding-box samples.

Fig. 4 shows the Precision-Recall (PR) curve with increasing β_1 angle thresholds (0 to π) for CLIP-hyp† and CLIP-euc†, and with cosine similarity thresholds (0 to 1) for the baseline CLIP. The proposed hyperbolic model trained without cross-image samples substantially outperforms the Euclidean model and even exceeds the Euclidean model with cross-image samples, indicating it implicitly learns the underlying structure. When cross-image samples are introduced, the hyperbolic model still leads, though the performance gap narrows.

Figure 5. **Example of parent-to-child retrieval using CLIP ViT and our CLIP-hyp† model.** Results are ordered by ascending norms. Our model retrieves images matching the predefined *scene-object-part* hierarchy, placing high-level objects near the origin (*e.g.*, harbor → boart parts), and grouping semantically related but visually distinct objects (*e.g.*, microwave oven & kitchen hood).

Qualitative Results. Following [6, 10, 32], we use parent-to-child image traversals with results ordered by increasing embedding norm, to illustrate the latent space, which: (1) aligns with the predefined *scene-object-part* hierarchy, placing high-level objects near the origin. (2). groups diverse objects under the same lower hierarchical branch, even if they are visually distinct (*e.g.,* keyboard and monitor under the studio image). Moreover, Figure 6 shows an example of the model's emergent structure where high-frequency, ambiguous image crops are naturally pushed toward the boundary of hyperbolic space. Find details in supplementary materials.

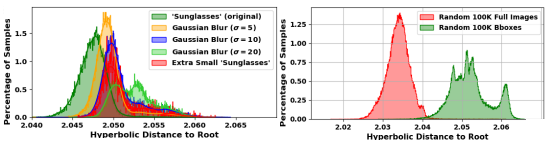


Figure 6. **Example of the model's emergent structure.**

5.2. Generalization Evaluation

We evaluate the generalization of our method through image retrieval on unseen datasets and hierarchies. Similar to Table 1-2, Table 3 presents part-based same-class and hierarchical retrieval performance on the **out-of-domain** LVIS dataset [13], which contains 1,203 fine-grained object classes, including many rare objects absent from the training *HierOpenImages* dataset. Table 4 evaluates image retrieval performance on VOC and COCO datasets following the zero-shot object detection in [29] where ground-truth boxes are used as region proposals, and then predict via top-5 majority vote. Results demonstrate that our visual hierarchical learning significantly enhances model generalization on unseen datasets and image hierarchies.

Model	Dist. Func.	Top-5	Top-10	Top-50	Top-100
Child-to-Parent					
CLIP	Cos Sim.	2.37	2.32	2.04	1.84
CLIP-euc†	Euc Ang.	19.36	18.64	16.14	14.52
CLIP-hyp†	Hyp Ang.	**22.33**	**21.40**	**18.50**	**16.42**
Parent-to-Child					
CLIP	Cos Sim.	8.57	7.77	5.70	4.64
CLIP-euc†	Euc Ang.	20.24	20.03	19.30	18.68
CLIP-hyp†	Hyp Ang.	**22.41**	**22.29**	**21.70**	**21.10**

Metrics	Model	Dist. Func.	Top-10k	Top-20k	Top-30k
Recall ↑	CLIP	Cos Sim.	18.30	32.42	45.42
	CLIP-euc†	Euc Ang.	46.89	64.40	75.71
	CLIP-hyp†	Hyp Ang.	**47.57**	**64.71**	**75.72**
OT Distance ↓	CLIP	Cos Sim.	7.61	8.88	9.43
	CLIP-euc†	Euc Ang.	5.74	7.39	8.43
	CLIP-hyp†	Hyp Ang.	**5.74**	**7.38**	**8.40**

Table 3. **Out-of-domain part-based image retrieval evaluation on the LVIS dataset:** same-class (top) and part-based hierarchical retrieval (bottom).

Dataset	CLIP	CLIP-euc†	CLIP-hyp†
VOC	87.3	92.2	**93.9**
COCO	63.6	66.5	**71.8**

Table 4. **Zero-shot object detection using ground-truth boxes.** Image-only model with top-5 majority voting.

6. Conclusion

In this work, we introduce a new learning paradigm that effectively encodes user-defined visual hierarchies in hyperbolic space without requiring explicit hierarchical labels. We present a concrete example of defining a part-based multi-level complex image hierarchy using object-level annotations and propose a contrastive loss in hyperbolic space to enforce pairwise entailment relationships. Additionally, we introduce new evaluation metrics for hierarchical image retrieval. Our experiments demonstrate our model effectively learns the predefined image hierarchy and goes beyond visual similarity.

References

[1] Morris Alper and Hadar Averbuch-Elor. Emergent visual-semantic hierarchies in image-text representations. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2024. 5, 6

[2] Martin R Bridson and André Haefliger. *Metric spaces of non-positive curvature*. Springer Science & Business Media, 2013. 2, 5

[3] Jiaxin Chen, Jie Qin, Yuming Shen, Li Liu, Fan Zhu, and Ling Shao. Learning attentive and hierarchical representations for 3d shape recognition. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XV 16*, pages 105–122. Springer, 2020. 5

[4] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *International conference on machine learning*, pages 1597–1607. PMLR, 2020. 1

[5] Xinlei Chen, Haoqi Fan, Ross Girshick, and Kaiming He. Improved baselines with momentum contrastive learning. *arXiv preprint arXiv:2003.04297*, 2020. 6

[6] Karan Desai, Maximilian Nickel, Tanmay Rajpurohit, Justin Johnson, and Shanmukha Ramakrishna Vedantam. Hyperbolic image-text representations. In *Internationaal Conference on Machine Learning*, pages 7694–7731. PMLR, 2023. 2, 4, 5, 6, 8

[7] Ankit Dhall, Anastasia Makarova, Octavian Ganea, Dario Pavllo, Michael Greeff, and Andreas Krause. Hierarchical image classification using entailment cone embeddings. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition workshops*, pages 836–837, 2020. 6

[8] Octavian Ganea, Gary Bécigneul, and Thomas Hofmann. Hyperbolic entailment cones for learning hierarchical embeddings. In *International Conference on Machine Learning*, pages 1646–1655. PMLR, 2018. 4

[9] Octavian Ganea, Gary Bécigneul, and Thomas Hofmann. Hyperbolic neural networks. *Advances in neural information processing systems*, 31, 2018. 5, 6

[10] Songwei Ge, Shlok Kumar Mishra, Haohan Wang, Chun-Liang Li, and David Jacobs. Robust contrastive learning using negative samples with diminished semantics. In *NeurIPS*, 2021. 8

[11] Songwei Ge, Shlok Mishra, Simon Kornblith, Chun-Liang Li, and David Jacobs. Hyperbolic contrastive learning for visual representations beyond objects. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6840–6849, 2023. 2, 5, 6, 7

[12] Yunhui Guo, Xudong Wang, Yubei Chen, and Stella X Yu. Clipped hyperbolic classifiers are super-hyperbolic classifiers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11–20, 2022. 5

[13] Agrim Gupta, Piotr Dollar, and Ross Girshick. Lvis: A dataset for large vocabulary instance segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 5356–5364, 2019. 6, 8

[14] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016. 1

[15] Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. Momentum contrast for unsupervised visual representation learning. In *CVPR*, 2020.

[16] Elad Hoffer and Nir Ailon. Deep metric learning using triplet network. In *Similarity-based pattern recognition: third international workshop, SIMBAD 2015, Copenhagen, Denmark, October 12-14, 2015. Proceedings 3*, pages 84–92. Springer, 2015. 1

[17] Joy Hsu, Jeffrey Gu, Gong Wu, Wah Chiu, and Serena Yeung. Capturing implicit hierarchical structure in 3d biomedical images with self-supervised hyperbolic representations. *Advances in neural information processing systems*, 34:5112–5123, 2021. 5

[18] Yoo Hyun Jeong, Myeongsoo Han, and Dong-Kyu Chae. A simple angle-based approach for contrastive learning of unsupervised sentence representation. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 5553–5572, 2024. 5

[19] Salman Khan, Muzammal Naseer, Munawar Hayat, Syed Waqas Zamir, Fahad Shahbaz Khan, and Mubarak Shah. Transformers in vision: A survey. *ACM computing surveys (CSUR)*, 54(10s):1–41, 2022. 5

[20] 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 Pat-*

tern Recognition, pages 6418–6428, 2020. 2, 3, 5, 6

[21] Sungyeon Kim, Boseung Jeong, and Suha Kwak. Hier: Metric learning beyond class labels via hierarchical regularization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 19903–19912, 2023. 5

[22] Fanjie Kong, Yanbei Chen, Jiarui Cai, and Davide Modolo. Hyperbolic learning with synthetic captions for open-world detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16762–16771, 2024. 5

[23] Alina Kuznetsova, Hassan Rom, Neil Alldrin, Jasper Uijlings, Ivan Krasin, Jordi Pont-Tuset, Shahab Kamali, Stefan Popov, Matteo Malloci, Alexander Kolesnikov, et al. The open images dataset v4: Unified image classification, object detection, and visual relationship detection at scale. *International journal of computer vision*, 128(7):1956–1981, 2020. 2, 6

[24] Qi Liu, Maximilian Nickel, and Douwe Kiela. Hyperbolic graph neural networks. *NeurIPS*, 2019. 5

[25] 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 *CVPR*, 2020. 6

[26] Maximillian Nickel and Douwe Kiela. Poincaré embeddings for learning hierarchical representations. *Advances in neural information processing systems*, 30, 2017. 1, 2, 5

[27] Maximillian Nickel and Douwe Kiela. Learning continuous hierarchies in the lorentz model of hyperbolic geometry. In *International conference on machine learning*, pages 3779–3788. PMLR, 2018. 2, 5

[28] Aaron van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*, 2018. 4

[29] 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. 6, 8

[30] Zexuan Qiu, Jiahong Liu, Yankai Chen, and Irwin King. Hihpq: Hierarchical hyperbolic product quantization for unsupervised image retrieval. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 4614–4622, 2024. 2, 5

[31] 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. 6

[32] Sameera Ramasinghe, Violetta Shevchenko, Gil Avraham, and Ajanthan Thalaiyasingam. Accept the modality gap: An exploration in the hyperbolic space. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 27263–27272, 2024. 2, 3, 4, 5, 6, 8

[33] John G Ratcliffe, S Axler, and KA Ribet. *Foundations of hyperbolic manifolds*. Springer, 1994. 2

[34] Erzsébet Ravasz and Albert-László Barabási. Hierarchical organization in complex networks. *Physical review E*, 67(2):026112, 2003. 1

[35] Frederic Sala, Chris De Sa, Albert Gu, and Christopher Ré. Representation tradeoffs for hyperbolic embeddings. In *ICML*, 2018. 2

[36] Florian Schroff, Dmitry Kalenichenko, and James Philbin. Facenet: A unified embedding for face recognition and clustering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 815–823, 2015. 1

[37] Alexandru Tifrea, Gary Bécigneul, and Octavian-Eugen Ganea. Poincar\'e glove: Hyperbolic word embeddings. *arXiv preprint arXiv:1810.06546*, 2018. 5

[38] Luke Vilnis, Xiang Li, Shikhar Murty, and Andrew McCallum. Probabilistic embedding of knowledge graphs with box lattice measures. *arXiv preprint arXiv:1805.06627*, 2018. 5

[39] Sijie Wang, Qiyu Kang, Rui She, Wei Wang, Kai Zhao, Yang Song, and Wee Peng Tay. Hypliloc: Towards effective lidar pose regression with hyperbolic fusion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5176–5185, 2023. 5

[40] Jiexi Yan, Lei Luo, Cheng Deng, and Heng Huang. Unsupervised hyperbolic metric learning. In *CVPR*, 2021. 5

[41] Zhen Yu, Toan Nguyen, Yaniv Gal, Lie Ju, Shekhar S Chandra, Lei Zhang, Paul Bonnington, Victoria Mar, Zhiyong Wang, and Zongyuan Ge. Skin lesion recognition with class-hierarchy regularized hyperbolic embeddings. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 594–603. Springer, 2022. 5

[42] Yudong Zhu, Di Zhou, Jinghui Xiao, Xin Jiang, Xiao Chen, and Qun Liu. Hypertext: Endowing fasttext with hyperbolic geometry. *arXiv preprint arXiv:2010.16143*, 2020. 5