HierVision: Standardized and Reproducible Hierarchical Sources for Vision Datasets

Anonymous ICCVW submission

Paper ID 10

Abstract

One-vs-rest training is a pervasive optimization regime in deep learning, whether the problem is supervised, self-supervised, or multi-modal in nature. The real world is however not binary, but governed by hierarchies. Hierarchies provide key information about the semantic relation between concepts, about which mistakes to avoid, and about the inherent organization of vision and language itself. Hierarchical learning, therefore, has a long history in computer vision and has gained further traction with the rise of hyperbolic deep learning. Currently, however, hierarchies are not standardized and centrally organized. Instead, such knowledge is scattered around various repositories, with inconsistent formatting, organizations, and availability. The lack of a central hub for hierarchies in vision datasets harms the utility and reproducibility of hierarchical learning. This paper introduces HierVision*, a central hub for hierarchical knowledge in vision datasets. This hub contains 60+ hierarchical sources, spanning actions, concepts, fine-grained categories, vision-language, and more. We outline a uniform coding of the hierarchies and procedures to embed them in existing pipelines. With this hub, we hope to positively impact the broad use and re-use of hierarchies for deep learning in computer vision.*

1. Introduction

Hierarchies are ubiquitous data structures across all sciences; from lesion taxonomies in the medical domain to animal ontologies in biology and semantic trees in natural language [2, 82]. Such tree-like structures have been used for centuries to organize our data and natural phenomena [1]. Computer vision deals with categorizing concepts from the real world, and datasets are therefore commonly organized hierarchically. Consider for example the WordNet hierarchy behind ImageNet [26], the biological ontology of birds in CUB [117], or the tree of verbs in Kinetics [58].

Despite the widespread availability of hierarchical information for computer vision, such knowledge is typically ignored when training deep networks. Instead, one-versus-rest optimization through cross-entropy and contrastive objectives are default options [21, 50]. Such a setup presents a binary view to categorization, where classes are either positive or (equally) negative. As a result the standard deep learning setup misses crucial hierarchical information about class similarities. The lack of hierarchical usage negatively impacts learning [53, 93, 126], generalization [93], error severity [8], and more.

An important reason for the lack of hierarchical integration in modern deep learning for vision is a geometric mismatch. Deep learning is Euclidean by default. Hierarchies are, however, exponentially growing structures, which leads to distortion when embedding them in Euclidean space [103], as Euclidean volumes grow only polynomially with their radius [88]. Recently, hyperbolic learning has rapidly gained traction in computer vision [80], as a natural space for embedding hierarchies [36, 88, 101, 115] and therefore a natural solution for hierarchical computer vision [3, 5, 28, 57, 74]. As a result, there is a growing demand for hierarchical knowledge in vision datasets.

An important issue currently is that there is no central hub for storing and sharing hierarchies. This does not align with the best scientific practices and hampers research. Not only are hierarchies arbitrarily hard to find depending on the dataset, but they are also not standardized and can even be altered. As such, it is unnecessarily hard to use hierarchies, and reproducibility is low since it is unknown whether hierarchies are identical. This paper introduces *HierVision*, a central hub for sharing hierarchies in vision datasets. Our goal is simple: create a continuous effort to store hierarchies for all vision datasets in a single place. Each hierarchy is standardized in a single format for ease of use and reproducibility. The hub also contains pipelines for visualization, analysis, and integration in deep learning and hyperbolic embedding pipelines. With *HierVision*, we want to make the community aware of the broad potential of hierarchical knowledge and the need for a central hub to organized computer vision hierarchically.

2. Related Work

2.1. Hierarchical Datasets

For clarity, we categorize prominent dataset hierarchies into two groups–those emphasizing semantic ontologies and those following biological taxonomies.

Semantic hierarchies are typically derived from human-defined knowledge bases or lexical resources. A foundational example is the use of WordNet [82], a large lexical database of English, to structure the ImageNet dataset [26]. This provided a rich, structured ontology that has been instrumental in the development of deep learning models. Other prominent datasets include CIFAR-100 [61], PASCAL-VOC [31], and OpenImages [62]. The BREEDS benchmark, derived from ImageNet, explicitly uses the class hierarchy to study robustness [102]. Such semantic structures are not limited to object recognition and extend to domains like medical imaging with datasets like CheXpert [109] and scene understanding with ADE20K [133].

The second category of datasets follows formal biological taxonomies, providing a scientifically grounded structure for fine-grained visual categorization. These datasets are critical for applications in biodiversity and conservation. For example, iNaturalist [114] organizes species observations according to the taxonomic rank (kingdom, phylum, class, etc.), ensuring that classes have a hierarchical relationship. TreeOfLife-10M and Rare Species [110], Classic fine-grained benchmarks CUB [117] and NABirds [113] are built on the taxonomy of species and genera, and datasets like AutoArborist [7] structure tree images by botanical taxonomy.

2.2. Hierarchies enhance vision tasks

The use of hierarchies as a source of prior knowledge is a long-standing concept in computer vision [26, 66, 78]. Classical approaches from the pre-deep learning era explicitly modeled the compositional nature of objects and scenes. Part-based approaches, such as pictorial structures [35] and grammar-based models [134], organize objects and scenes into parts to improve image recognition and interpretability. [34, 35, 75, 107, 134].

In the modern deep learning era, hierarchical knowledge is integrated through various mechanisms such as class taxonomies, structured loss functions, and specialized architectures [8, 9, 38, 85]. These approaches improve many tasks such including image classification [19, 53, 93], action classification [43, 74], and robustness to distribution shifts [102]. Hierarchical approaches were also used to measure the severity of classification mistakes, where misclassifying an object as a close relative in the hierarchy is penalized less severely [8, 9, 38, 39]. Furthermore, hierarchical information has been successfully incorporated into contrastive learning [12, 13, 46], and vision-language models [37, 90]. The utility of hierarchical methods also extends to applied domains such as medical imaging [18] and autonomous driving [83]. A common theme of these works is their reliance on Euclidean geometry to model these hierarchical relationships.

2.3. Hyperbolic learning

Hyperbolic learning has emerged as a powerful paradigm for encoding and exploiting hierarchical relationships in visual data. Owing to the constant negative curvature of hyperbolic space, it can be thought of as a continuous version of a tree, making it a good choice to accommodates tree-like structures while preserving distances [48, 112]. In recent years, hierarchical embeddings have been performed in hyperbolic space [88], leading to successfully embedding complex trees with low distortions [36, 63, 89, 101, 115, 129].

Many computer vision tasks inherently involve hierarchies, for example, semantic grouping or biological taxonomies (Sec. 2.1). A wide range of works have recently shown the potential and effectiveness of using a hyperbolic embedding space for both supervised and unsupervised learning [80]. Specifically, in supervised settings, the hierarchical prior knowledge of the datasets can be embedded in hyperbolic space, after which the visual representations can be mapped to the same space and optimized to match this hierarchical organization [5, 57, 74]. Hyperbolic embeddings have shown benefits in classification [41, 45, 116], segmentation [3, 17], out-of-distribution detection [36, 57, 116], uncertainty quantification [3, 17], zero-shot learning [4, 51, 69], continual learning [5, 24, 111], hierarchical representation learning [28, 29, 72, 74], contrastive learning [40, 130], generative models [64, 95], and vision-language models [27, 54, 92, 94]. Recently Ayoughi *et al*. [6] discussed optimial tree structure for hyperbolic embeddings.

3. Hierarchies

While the benefits of hierarchical information in computer vision are well-established, its practical adoption has been limited by fragmented and inconsistent hierarchy management across datasets. Hierarchies exist in various formats, from simple text files and custom XML/JSON to folder-based organizations, which require custom parsing for each data set. This fragmentation hinders reproducibility and the development of hierarchical computer vision methods. To address this, we introduce *HierVision*, a centralized hub that standardizes hierarchy representation and enables seamless integration with existing tools. In the following sections, we first define a common hierarchy format that can be used with graph processing library like NetworkX [47] (Section 3.1). Then we briefly discuss describe our standardization process (Section 3.2. We then discuss in detail the cur-

rent datasets (Section 3.3) and finally present some dataset hierarchy statistics (Section 3.4).

3.1. Hierarchy format

To enable standardized storage and use of hierarchical information across diverse vision datasets, we adopt a uniform graph-based representation format. Each hierarchy is modeled as a directed, rooted tree encoded in JSON. Formally, we represent a hierarchy as a graph, $G = (V, E)$, where V is the set of nodes (*e.g.* dataset classes or parents) and $E \in V \times V$ is the set of directed edges, such that an edge $(u, v) \in E$ denotes a parent-child relationship and node v is a subclass of node u. In practice, each hierarchy is stored in a JSON with the following components:

- `"nodes"`: A list of nodes, each with an integer `"id"` and a human-readable `"label"`. These define the concepts in the hierarchy.
- `"links"`: A list of directed edges, each represented as a with `"source"` and `"target"` keys indicating parent and child node IDs, respectively.
- `"directed"`: A boolean flag, always set to `true`.
- `"multigraph"`: A boolean flag, always set to `false`, enforcing at most one edge between any pair of nodes.

An excerpt of a simplified hierarchy is shown below:

```json
{
  "directed": true,
  "multigraph": false,
  "nodes": [
    {"id": 3, "label": "root"},
    {"id": 2, "label": "animal"},
    {"id": 1, "label": "dog"},
    {"id": 0, "label": "cat"}
  ],
  "links": [
    {"source": 3, "target": 2},
    {"source": 2, "target": 0},
    {"source": 2, "target": 1}
  ]
}
```

Listing 1. A sample JSON format of simple hierarchy. All hierarchies in *Hiervision* follow this structure, allowing easy visualization and use in downstream applications

In this example, `"root"` is the global ancestor of all nodes, `"animal"` is a direct child of `"root"`, and both `"cat"` and `"dog"` are subclasses of `"animal"`. This structure is fully compatible with standard graph-processing libraries, particularly NetworkX [47], which we use throughout our implementation. Hierarchies can be directly loaded as `networkx.DiGraph` objects, enabling efficient hierarchy traversal, visualization, validation, and integration into hierarchical deep learning pipelines, including those that rely on hyperbolic embeddings or hierarchy-aware loss functions.

3.2. Standardizing the dataset hierarchies

We collect hierarchies from over 61 vision datasets spanning various domains such as object recognition, fine-grained classification, action understanding, scene interpretation, video analysis, and medical imaging. These hierarchies originate from either the dataset creators themselves or other papers that construct or refine hierarchies of existing datasets for specific tasks. Across these sources, we observe significant variation in format: some hierarchies are in graph-structured files (e.g., JSON, XML trees). Others use flat lists with indentation or prefix-based identifiers to imply structure. Some are only documented visually or embedded in figures or tables in papers. These sources vary significantly in format, structure, and accessibility, requiring a standardization process.

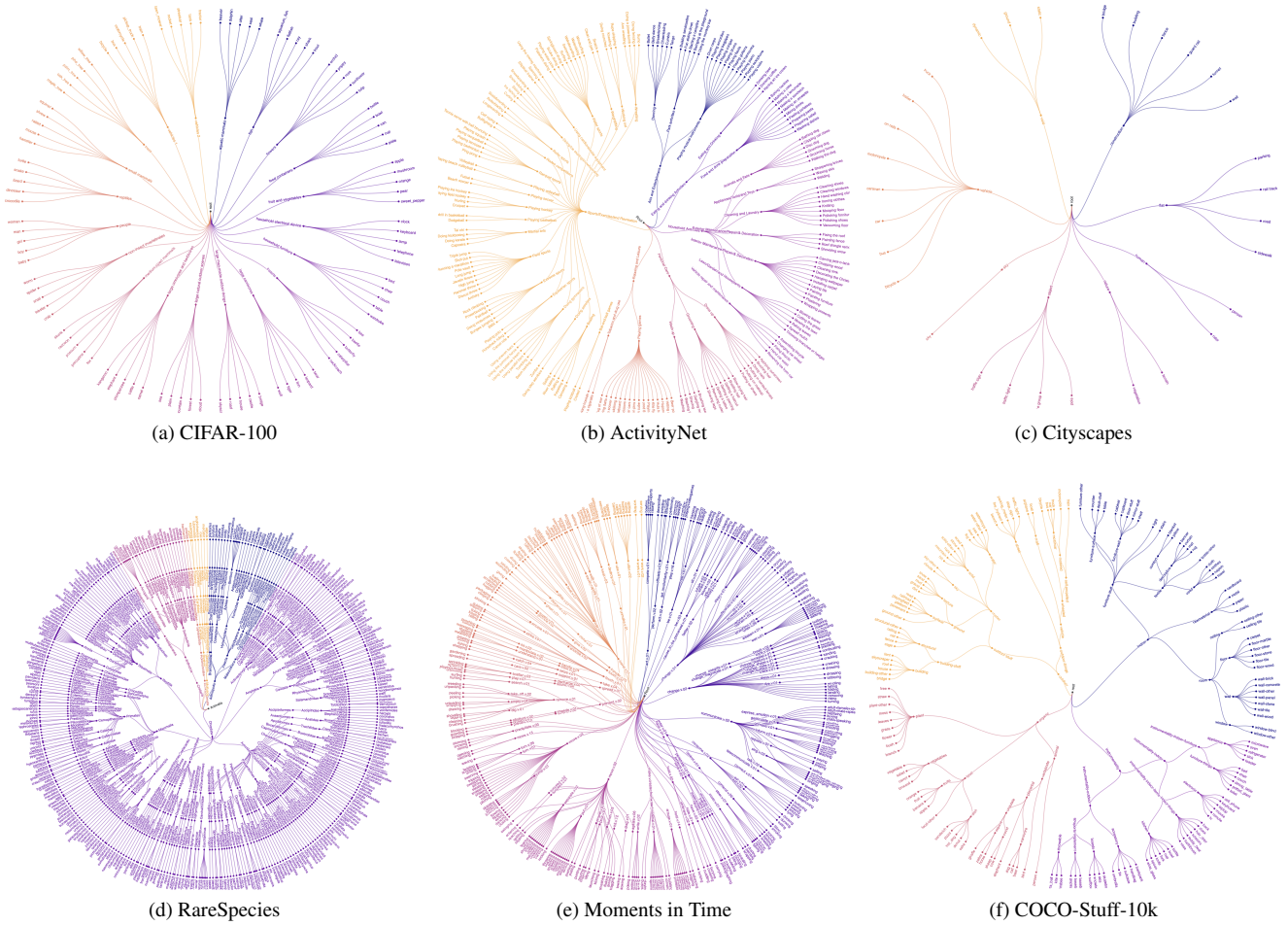
To unify these into a standardized format, we add each hierarchy into the JSON graph structure described in Sec. 3.1. We parse and extract the node and edge structure. We resolve any label inconsistences and add a single root node. We validate the resulting graph using NetworkX to ensure it forms a connect, directed tree.

In cases where multiple versions of a hierarchy exist (e.g., fine vs. coarse levels), we store each version explicitly. This ensures that downstream tasks can choose the level of abstraction best suited to their needs.

3.3. Dataset coverage

Our *HierVision* hub currently covers over 50 vision datasets spanning a wide range of domains and recognition tasks. The collection includes image classification datasets such as CIFAR-100, ImageNet-100, ImageNet-1K, ImageNet-21K, and iNaturalist; semantic segmentation datasets including ADE20K and Cityscapes; action recognition datasets such as Kinetics, ActivityNet, and FineSports; fine-grained categorization with CUB, FGVC-Aircraft, and TreeOfLife-10M; and medical imaging datasets like CheXpert and DeepLesion. Hierarchies in these datasets range from shallow groupings of a few categories to deeply nested structures with thousands of classes, reflecting both semantic and biological taxonomies. Further details, including taxonomy type and hierarchy source, are summarized in Table 1.

Figure 1 shows visualizations of the hierarchical structures for six representative datasets: CIFAR-100 (semantic object classification with a two-level hierarchy), Cityscapes (urban scene segmentation with a class grouping tree), RareSpecies (a deeply nested biological taxonomy), ActivityNet (action recognition with a hierarchical verb ontology), Moments in Time (an action dataset with a flat class structure), and COCO-Stuff-10k (scene parsing with a multi-level segmentation hierarchy). These examples highlight the diversity in both the structure and scale of the hierarchies, ranging from compact, balanced trees to large, irregular graphs with hundreds or thousands of nodes.



(a) CIFAR-100 (b) ActivityNet (c) Cityscapes

(d) RareSpecies (e) Moments in Time (f) COCO-Stuff-10k

Figure 1. Hierarchy visualizations for a selection of datasets form *HierVision*. The coverage in the repository hub ranges from small number of classes and shallow trees to very large number of classes and deep trees.

3.4. Hierarchy Statistics

We visualize statistics across all collected hierarchies to assess their structural diversity. Figure 2 summarizes four key aspects: node count, maximum depth, the relationship between node count and depth, and average branching factor.

Most hierarchies contain fewer than 1,000 classes, though there is a long tail of large-scale taxonomies such as TreeOfLife-10M [110], ImageNet-21K [98], and Bamboo [132], that reach tens or hundreds of thousands of nodes (Figure 2a).

The majority of hierarchies are shallow, with depths of 2 or 3, as seen in datasets like CIFAR-100 and ADE20K (Figure 2b). However, a subset—including TreeOfLife-10M [110] and Visual Genome [60]—features deeply nested structures, with depths exceeding 10. The scatter plot of node count versus maximum depth (Figure 2c) highlights this diversity: some datasets combine high node counts and depth, while others are both small and shallow.

Average branching factor calculates the average number of children for each node, which also varies widely (Figure 2d). Biological taxonomies tend to be deep and balanced, with low branching factors (1–3), while semantic hierarchies such as ADE20K [133] and Objects-365 [104] are broad and flat, with extremely high branching at the root.

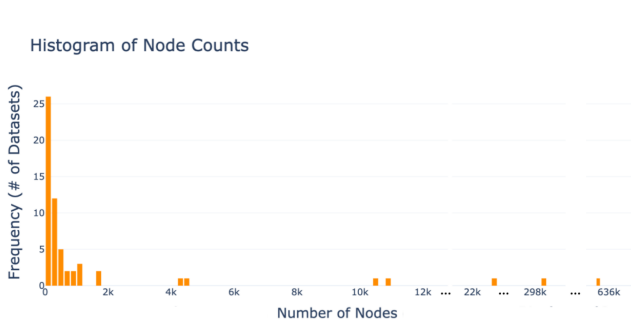
4. Integration into Deep Learning Pipelines

The standardized graph-based format adopted by *HierVision* enables seamless integration of hierarchical information into modern deep learning workflows. In this section, we outline typical approaches for incorporating hierarchies, ranging from data loading and label preprocessing to advanced hierarchical loss design and representation learning.

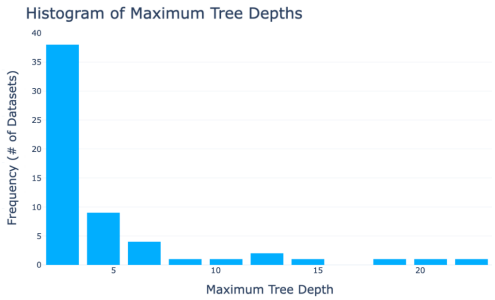
Hierarchies stored in our JSON format can be loaded directly using widely adopted graph libraries such as NetworkX:

Table 1. All hierarchies available currently in *HierVision*. Datasets with multiple hierarchy versions (e.g., coarse/fine) are marked with *. If the hierarchy was sourced from another paper, it is cited in the "Hierarchy Source" column.

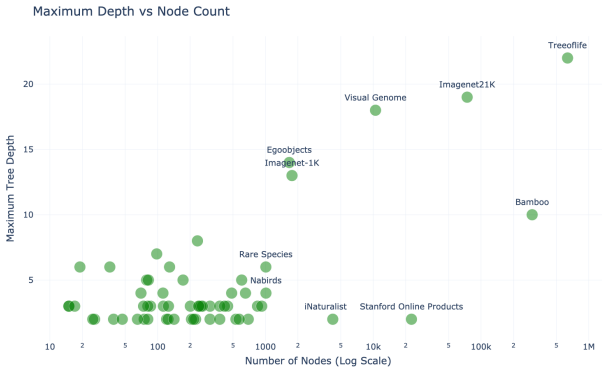
	Dataset	Hierarchy Source	Original Format	Nodes	Edges	Depth	Classes
Actions / Video	ActivityNet [14]	-	JSON	245	244	3	200
	DLD3V-10K [68]	-	JSON	81	80	2	64
	FineSports [124]	-	PKL	65	64	2	52
	HDM05 [86]	-	JSON	26	25	2	20
	HowTo100M [81]	-	JSON	142	141	2	129
	HumanAct12 [44]	-	NPY	47	46	2	34
	MAdVerse [100]	-	JSON	653	655	4	578
	Matador [10]	-	WEB	82	81	5	59
	Mini-Kinetics-200 [58, 123]	-	JSON	240	239	3	200
	MIntRec* [131]	-	TSV	25	24	2	22
	Moments in Time [84]	-	JSON	486	485	4	339
	Pseudo-Adverbs (ActivityNet) [30]	-	CSV	758	757	2	643
	Pseudo-Adverbs (MSRVTT) [30]	-	CSV	571	570	2	464
	Pseudo-Adverbs (VATEX) [30]	-	CSV	1686	1685	2	1550
	Something-Something V2* [42, 76]	[76]	JSON	225	224	2	174
	UCF101 [108]	-	CSV	127	126	3	100
Biological	AwA2 [120]	-	TXT	86	85	3	50
	BioTrove-Balanced [127]	-	CSV	818	823	7	292
	BioTrove-LifeStages [127]	-	CSV	19	18	6	5
	BioTrove-Unseen [127]	-	CSV	4413	4703	7	1918
	CUB-200-2011 [117]	-	JSON	251	250	3	200
	iNaturalist [114]	-	CSV	4214	4213	2	4200
	MammalNet [20]	-	TSV	260	264	3	173
	Marine Tree [11]	-	CSV	79	78	5	62
	NABirds [113]	-	TXT	1011	1010	4	555
	Rare Species [110]	-	CSV	1009	1023	6	385
	Tree of Life [110]	-	CSV	635463	635462	22	537235
	VegFru-Fru92 [52]	-	JSON	103	102	2	92
	VegFru-Veg200 [52]	-	JSON	216	215	2	200
Medical	CheXpert [55]	-	CSV	15	15	2	11
	DeepLesion [125]	-	CSV	172	273	2	117
	MIMIC-CXR [55]	-	CSV	15	15	2	11
	OpenCell [22]	-	CSV	17	16	3	11
Objects / General	Bamboo [132]	-	JSON	298307	245557	2	285340
	Caltech-101 [65]	-	Folder	113	113	4	102
	CIFAR-100 [61]	-	PKL	121	120	2	100
	CMU MoCap [25]	-	WEB	306	305	3	280
	COCO-10K [15]	[3]	JSON	234	233	8	171
	COD10K [32]	-	JSON	75	74	2	69
	CORe50-Balanced [71]	-	TXT	70	69	4	50
	CORe50-Unbalanced [71]	-	TXT	66	65	5	50
	EgoObjects [33]	-	JSON	1665	1664	14	1179
	FGVC-Aircraft [77]	-	TXT	201	200	3	100
	Fashionpedia-Attributes [56]	-	JSON	306	305	2	294
	Fashionpedia-Categories [56]	-	JSON	62	61	3	46
	IP102 [119]	-	TXT	113	112	3	102
	ImageNet-100* [26]	[67]	Folder	121	120	2	100
	ImageNet-1K [26]	-	Folder	1763	1776	13	1010
	ImageNet-21K* [98]	-	Folder	10891	11734	18	7414
	ImageNet-OOD* [128]	[98]	Folder	954	967	13	630
	Objects365 [104]	-	WEB	377	376	2	365
	OpenLORIS [105]	-	JSON	98	97	7	69
	OpenImages [62]	-	JSON	603	648	5	525
	PASCAL VOC [31]	[3]	XML	36	35	6	21
	Portrait Mode 400 [49]	-	CSV	446	445	3	400
	Stanford Cars [59]	-	Folder	206	205	2	196
	Stanford Online Products [91]	-	Folder	22634	22633	2	22622
Scenes / Places	ADE20K* [133]	[3]	JSON	1116	1115	2	1105
	Cityscapes [23]	-	WEB	39	38	2	30
	Grocery [79]	-	CSV	125	124	2	81
	Mapillary Vistas [87]	-	JSON	81	80	3	66
	Million-AID [73]	-	XML	74	73	3	51
	PACO-EGO4D [96]	-	JSON	515	514	2	441
	PACO-LVIS [96]	-	JSON	532	531	2	458
	SUN360 [122]	-	WEB	378	377	3	356
	SUN397 [121]	-	CSV	418	544	3	397
	SUN908 [121]	-	WEB	925	1088	3	904
	Visual Genome [60]	-	JSON	10503	10502	18	6114



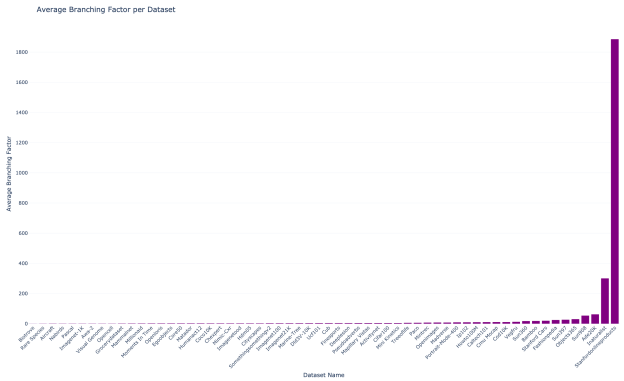
(a) Node Count Distribution



(b) Max Depth Distribution



(c) Node count vs Max Depth



(d) Average Branching Factor

Figure 2. Statistics and distributions of hierarchies in our *HierVision* collection.

```python
import json, networkx as nx
with open('hierarchy.json') as f:
    graph_dict = json.load(f)
G = nx.node_link_graph(graph_dict)
```

Listing 2. Loading a hierarchy into `networkx`.

Once loaded as a graph object, the hierarchy can be queried to obtain ancestor or descendant sets for each class, compute semantic distances between nodes, or extract sub-hierarchies for specialized tasks. Below, we discuss the relevance of hierarchies to deep learning in euclidean space (Section 4.1) and hyperbolic space (Section 4.2. Additionally in hyperbolic learning, we show how the few of the hierarchies can be embedded into hyperbolic space.

4.1. Relevance to Hierarchy-Aware Supervision

4.1.1. Hierarchy-Aware Label Representations and Multi-Level Outputs

HierVision enables augmenting dataset labels with hierarchical context by allowing straightforward retrieval of parent and ancestor labels for any fine-grained class. This facilitates multi-task learning, where models predict class probabilities at multiple hierarchical levels (e.g., object category and super-category) simultaneously, typically via auxiliary output heads and backpropagating loss at each level [118, 126]. Such coarse-to-fine supervision guides feature learning: high-level layers discern broad distinctions, while deeper layers refine for fine-grained classification.

Alternatively, hierarchical classification [106] predicts sequentially, predicting a coarse category before specializing to specific subclasses [16, 70, 93, 118]. In both approaches, *HierVision*'s standardized graph simplifies retrieving relevant ancestor or child classes, ensuring hierarchy-consistent predictions and providing interpretable outputs at multiple levels of detail.

4.1.2. Hierarchy-Aware Loss Functions

HierVision facilitates the design of loss functions that account for inter-class relationships, moving beyond standard cross-entropy's uniform error penalty. By leveraging the hierarchy, errors can be weighted by their distance in the tree or reflected in structured objectives.

A classic example is hierarchical softmax [85], which factors the prediction over a tree of classes. Instead of a flat N-class prediction, the model predicts a path from root to leaf, decomposing the problem into a sequence of smaller

binary classification tasks. For a tree with internal nodes $\mathcal{N}$, let y denote the true class and $\mathbf{z}$ the logits. The hierarchical softmax decomposes the probability of class y as:

$$P(y \mid \mathbf{z}) = \prod_{n \in \text{path}(y)} P(n \mid \text{parent}(n), \mathbf{z}) \quad (1)$$

where $\text{path}(y)$ is the sequence of nodes from the root to y. The loss is then computed as the negative log-probability along this path:

$$\mathcal{L}_{\text{hier}-\text{softmax}} = -\log P(y \mid \mathbf{z}) \quad (2)$$

This approach not only reduces computational cost for large N but also implicitly aligns internal representations with the taxonomy.

Beyond hierarchical softmax, cost-sensitive losses penalize mistakes according to taxonomy distance, e.g., based on the length of the shortest path or the depth of the lowest common ancestor (LCA) between the predicted class $\hat{y}$ and true class y [8]. Let $d(y, \hat{y})$ denote the tree distance between classes. The loss can be defined as:

$$\mathcal{L}_{\text{hier}-\text{LCA}} = \sum_{i=1}^{N} d(y_i, \hat{y}_i) \cdot \ell(y_i, \hat{y}_i) \quad (3)$$

where ℓ is the standard loss (e.g., cross-entropy), and $d(y, \hat{y})$ is typically the path length or a normalized form thereof.

Similarly, hierarchy-based label smoothing replaces one-hot targets with soft distributions, assigning higher probability mass to classes near the ground truth in the hierarchy [8, 97, 99]. For each target y, define the smoothed label $\tilde{y}$ as:

$$\tilde{y}_j = \frac{\exp(-\lambda \cdot d(y, j))}{\sum_{k=1}^{N} \exp(-\lambda \cdot d(y, k))} \quad (4)$$

where $d(y, j)$ is the distance in the hierarchy between y and j, and $\lambda > 0$ is a hyperparameter controlling the smoothing. The loss is then the standard cross-entropy between the prediction and $\tilde{y}$:

$$\mathcal{L}_{\text{hier}-\text{smooth}} = -\sum_{j=1}^{N} \tilde{y}_j \log p_j \quad (5)$$

where p_j is the predicted probability for class j.

These hierarchy-aware objectives guide models to make semantically meaningful predictions, improve error robustness, and support finer-grained evaluation of learning performance.

4.2. Hyperbolic Learning

Hyperbolic learning uses the properties of hyperbolic space to better represent hierarchical relationships in vision datasets. Unlike Euclidean space, hyperbolic space expands exponentially, making it ideal for embedding tree-like taxonomies with minimal distortion. In practice, both class labels and image features are embedded as points in a hyperbolic manifold (e.g., the Poincaré ball), so that distances between points correspond to semantic or taxonomic proximity in the hierarchy.

	Poincaré [88]		Entailment [36]		BHE [57]	
	Dist	mAP	Dist	mAP	Dist	mAP
CIFAR-100 [61]	0.713	0.162	0.18	0.623	0.026	0.885
ImageNet-100 [26]	0.450	0.119	0.20	65.91	0.095	0.746
CityScapes [23]	0.540	0.173	0.250	72.41	0.050	0.967
PASCAL-VOC [31]	0.477	0.122	0.182	0.692	0.05	0.837

Table 2. Hyperbolic embeddings of CIFAR-100, ImageNet-100, Cityscapes and PASCAL-VOC in using different hyperbolic embedding methods. Distortion and mAP [101] of the embeddings measure how well the tree distances are embedded in the hyperbolic space.

For hyperbolic learning with vision datasets, a typical process has two main steps. Step 1 is embedding the hierarchy itself: the class tree is mapped into hyperbolic space so that classes which are close in the hierarchy are also close together in the embedding. This is usually done using methods such as Poincaré embeddings [88] or entailment-based approaches [36], which are specifically designed to preserve the distances and relationships from the original tree. The effectiveness of this embedding is measured by two metrics following the hyperbolic learning literature [88, 101, 103]: distortion, which captures how well the hyperbolic distances match the true tree structure (lower is better), and mean average precision (mAP), which reflects how well nearest neighbors in the embedding correspond to true neighbors in the hierarchy (higher is better).

Step 2, as done in works like [5, 57, 74], involves learning: projecting image features produced by a neural network into hyperbolic space and training them to align with their respective class embeddings. This allows the model to make predictions that are consistent with the structure of the hierarchy.

In this section, we focus on Step 1 and analyze how well different hyperbolic embedding methods can represent the hierarchy itself in Table 2. Using standardized trees from HierVision, we evaluate and compare several approaches on CIFAR-100, ImageNet-1k, ActivityNet, and CUB. For each dataset, we report distortion and mAP scores to assess how faithfully the class relationships are captured in hyperbolic space. We use the hyperparameters defined in Kasarla *et al.* [57] for generating the hyperbolic embeddings of the methods, keeping hyperbolic $dim = 64$. In Table 2, BHE [57] show better faithful tree embeddings. However, for downstream tasks, any of these embeddings can be used depending on the utility for the task.

The standardized JSON hierarchies in our hub make it easy to plug in such methods – for example, one can directly use NetworkX to compute ancestor relations or to feed the graph into a Poincaré embedding algorithm to obtain initial class vectors. The result is an integrated hyperbolic pipeline where both the data and the output of the model are easily intergrated in the hyperbolic space.

5. Conclusion

We introduce *HierVision*, a hub for standardized hierarchical knowledge across a broad spectrum of visual recognition datasets. By consolidating and curating over 61 hierarchies from diverse domains and encoding them in a unified graph-based format, we provide consistent, reproducible access to structured label information. Our framework supports direct integration with graph libraries like NetworkX and enables hierarchical loss design, hyperbolic embeddings, and large-scale benchmarking.

We believe that *HierVision* serves as a critical resource for the vision community, promoting reproducibility, accelerating hierarchy-informed research, and enabling rigorous benchmarking across a wide range of hierarchical structures.

We will make the GitHub repo for *HierVision* public after the review period. We shortlisted 40+ more hierarchies in the pipeline to be added in the future. Recently, Ayoughi *et al*. [6] discussed optimial tree structure for hyperbolic embeddings, which can be further used to refine the existing hierarchies. We invite community contributions to further expand and refine *HierVision*, and we hope this hub will catalyze advances in hierarchical representation learning, benchmarking, and structured visual understanding at scale.

References

[1] Aristotle. *Categories*. Clarendon Press, Oxford, 350 BCE. Translated and edited by J.L. Ackrill, 1963. 1

[2] Michael Ashburner, Catherine A Ball, Judith A Blake, David Botstein, Heather Butler, J Michael Cherry, Allan P Davis, Kara Dolinski, Selina S Dwight, Janan T Eppig, et al. Gene ontology: tool for the unification of biology. *Nature genetics*, 25(1):25–29, 2000. 1

[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. 1, 2, 5

[4] Mina Ghadimi Atigh, Stephanie Nargang, Martin Keller-Ressel, and Pascal Mettes. Simzsl: Zero-shot learning beyond a pre-defined semantic embedding space. *International Journal of Computer Vision*, pages 1–17, 2025. 2

[5] Melika Ayoughi, Mina Ghadimi Atigh, Mohammad Mahdi Derakhshani, Cees GM Snoek, Pascal Mettes, and Paul Groth. Continual hyperbolic learning of instances and classes. *arXiv preprint arXiv:2506.10710*, 2025. 1, 2, 7

[6] Melika Ayoughi, Max van Spengler, Pascal Mettes, and Paul Groth. Designing hierarchies for optimal hyperbolic embedding. In *European Semantic Web Conference*, pages 362–382. Springer, 2025. 2, 8

[7] Sara Beery, Guanhang Wu, Trevor Edwards, Filip Pavetic, Bo Majewski, Shreyasee Mukherjee, Stanley Chan, John Morgan, Vivek Rathod, and Jonathan Huang. The auto arborist dataset: a large-scale benchmark for multiview urban forest monitoring under domain shift. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 21294–21307, 2022. 2

[8] Luca Bertinetto, João F Henriques, and Philip HS Torr. Making the most of mistake-making: Hierarchical classification with partial labels. In *European Conference on Computer Vision (ECCV)*, pages 706–722. Springer, 2020. 1, 2, 7

[9] Luca Bertinetto, Romain Mueller, Konstantinos Tertikas, Sina Samangooei, and Nicholas A Lord. Making better mistakes: Leveraging class hierarchies with deep networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12506–12515, 2020. 2

[10] M. Beveridge and S. K. Nayar. (h)ierarchical (m)aterial (r)ecognition (f)rom (l)ocal (a)ppearance, 2025. 5

[11] Thomas Boone et al. Marine-tree: Marine species classification. GitHub repository, 2021. 5

[12] Guy Bukchin, Eli Schwartz, Kate Saenko, Ori Shahar, Rogerio Feris, Raja Giryes, and Leonid Karlinsky. Fine-grained angular contrastive learning with coarse labels. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 8730–8740, 2021. 2

[13] Hadas Bukchin, Yuval Alaluf, and Daniel Cohen-Or. Fine-grained object recognition via deep contrastive learning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV) Workshops*, pages 1534–1543, 2021. 2

[14] Fabian Caba Heilbron, Victor Escorcia, Joao Carreira, Juan Carlos Niebles, and Bernard Ghanem. Activitynet: A large-scale video benchmark for human activity understanding. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 961–970, 2015. 5

[15] Holger Caesar, Jasper Uijlings, and Vittorio Ferrari. Cocostuff: Thing and stuff classes in context. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1209–1218, 2018. 5

[16] Dongliang Chang, Kaiyue Pang, Yixiao Zheng, Zhanyu Ma, Yi-Zhe Song, and Jun Guo. Your" flamingo" is my" bird": Fine-grained, or not. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11476–11485, 2021. 6

[17] Bike Chen, Wei Peng, Xiaofeng Cao, and Juha Röning. Hyperbolic uncertainty aware semantic segmentation. *IEEE Transactions on Intelligent Transportation Systems*, 25(2): 1275–1290, 2023. 2

[18] Haomin Chen, Shun Miao, Daguang Xu, Gregory D Hager, and Adam P Harrison. Deep hierarchical multi-label classification of chest x-ray images. In *International conference on medical imaging with deep learning*, pages 109–120. PMLR, 2019. 2

[19] Jingzhou Chen, Peng Wang, Jian Liu, and Yuntao Qian. Label relation graphs enhanced hierarchical residual network for hierarchical multi-granularity classification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4858–4867, 2022. 2

[20] Jun Chen, Ming Hu, Darren J Coker, Michael L Berumen, Blair Costelloe, Sara Beery, Anna Rohrbach, and Mohamed Elhoseiny. Mammalnet: A large-scale video benchmark for mammal recognition and behavior understanding. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 13052–13061, 2023. 5

[21] 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

[22] Nathan H Cho, Keith C Cheveralls, Andreas-David Brunner, Kibeom Kim, André C Michaelis, Preethi Raghavan, Hirofumi Kobayashi, Laura Savy, Jason Y Li, Hera Canaj, et al. Opencell: Endogenous tagging for the cartography of human cellular organization. *Science*, 375(6585):eabi6983, 2022. 5

[23] Marius Cordts, Mohamed Omran, Sebastian Ramos, Timo Rehfeld, Markus Enzweiler, Rodrigo Benenson, Uwe Franke, Stefan Roth, and Bernt Schiele. The cityscapes dataset for semantic urban scene understanding. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3213–3223, 2016. 5, 7

[24] Yawen Cui, Zitong Yu, Wei Peng, Qi Tian, and Li Liu. Rethinking few-shot class-incremental learning with open-set hypothesis in hyperbolic geometry. *IEEE Transactions on Multimedia*, 26:5897–5910, 2023. 2

[25] Fernando De la Torre, Jessica Hodgins, Adam Bargteil, Xavier Martin, Justin Macey, Alex Collado, and Pep Beltran. Guide to the carnegie mellon university multimodal activity (cmu-mmac) database. 2009. 5

[26] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pages 248–255. IEEE, 2009. 1, 2, 5, 7

[27] 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

[28] Ankit Dhall, Anastasia Makarova, Octavian Ganea, Dario Pavllo, Michael Greeff, and Andreas Krause. Hierarchical image classification using entailment cone embeddings. In *CVPR Workshop on Differential Geometry in Computer Vision and Machine Learning*, 2020. 1, 2

[29] Lars Doorenbos, Pablo Márquez-Neila, Raphael Sznitman, and Pascal Mettes. Hyperbolic random forests. *arXiv preprint arXiv:2308.13279*, 2023. 2

[30] Hazel Doughty and Cees G. M. Snoek. How Do You Do It? Fine-Grained Action Understanding with Pseudo-Adverbs. In *The IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022. 5

[31] Mark Everingham, Luc Van Gool, Christopher KI Williams, John Winn, and Andrew Zisserman. The pascal visual object classes (voc) challenge. *International Journal of Computer Vision*, 88(2):303–338, 2010. 2, 5, 7

[32] Deng-Ping Fan, Ge-Peng Ji, Guolei Sun, Ming-Ming Cheng, Jianbing Shen, and Ling Shao. Camouflaged object detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 2777–2787, 2020. 5

[33] Jiyang Fan, Bo Xiong, Lu Jiang, Yin Cui, Chen Sun, Manmohan Chandraker Jain, and Kristen Grauman. Egoobjects: A large-scale egocentric dataset for object understanding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021. 5

[34] Clement Farabet, Camille Couprie, Laurent Najman, and Yann LeCun. Learning hierarchical features for scene labeling. *IEEE transactions on pattern analysis and machine intelligence*, 35(8):1915–1929, 2012. 2

[35] Pedro F Felzenszwalb, Ross B Girshick, and David McAllester. Cascade object detection with deformable part models. In *2010 IEEE Computer society conference on computer vision and pattern recognition*, pages 2241–2248. Ieee, 2010. 2

[36] 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. 1, 2, 7

[37] Yuting Gao, Jinfeng Liu, Zihan Xu, Jun Zhang, Ke Li, Rongrong Ji, and Chunhua Shen. Pyramidclip: Hierarchical feature alignment for vision-language model pretraining. *Advances in neural information processing systems*, 35:35959–35970, 2022. 2

[38] Ashima Garg, Depanshu Sani, and Saket Anand. Learning hierarchy aware features for reducing mistake severity. In *European Conference on Computer Vision*, pages 252–267. Springer, 2022. 2

[39] Ashima Garg, Depanshu Sani, and Saket Anand. Learning hierarchy aware features for reducing mistake severity. In *European Conference on Computer Vision*, pages 252–267. Springer, 2022. 2

[40] 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

[41] Mina Ghadimi Atigh, Martin Keller-Ressel, and Pascal Mettes. Hyperbolic busemann learning with ideal prototypes. *Advances in neural information processing systems*, 34:103–115, 2021. 2

[42] Raghav Goyal, Samira Ebrahimi Kahou, Vincent Michalski, Joanna Materzynska, Susanne Westphal, Heuna Kim, Valentin Haenel, Ingo Fruend, Peter Yianilos, Moritz Mueller-Freitag, et al. The" something something" video

database for learning and evaluating visual common sense. In *Proceedings of the IEEE international conference on computer vision*, pages 5842–5850, 2017. 5

[43] Sadaf Gulshad, Teng Long, and Nanne van Noord. Hierarchical explanations for video action recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3703–3708, 2023. 2

[44] Chuan Guo, Xinxin Zuo, Sen Wang, Shihao Zou, Qingyao Sun, Annan Deng, Minglun Gong, and Li Cheng. Action2motion: Conditioned generation of 3d human motions. In *Proceedings of the 28th ACM International Conference on Multimedia*, pages 2021–2029, 2020. 5

[45] 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. 2

[46] Yuanfan Guo, Minghao Xu, Jiawen Li, Bingbing Ni, Xuanyu Zhu, Zhenbang Sun, and Yi Xu. Hcsc: Hierarchical contrastive selective coding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9706–9715, 2022. 2

[47] Aric Hagberg, Pieter J Swart, and Daniel A Schult. Exploring network structure, dynamics, and function using networkx. Technical report, Los Alamos National Laboratory (LANL), Los Alamos, NM (United States), 2008. 2, 3

[48] Matthias Hamann. On the tree-likeness of hyperbolic spaces. In *Mathematical proceedings of the cambridge philosophical society*, pages 345–361. Cambridge University Press, 2018. 2

[49] Mingfei Han, Linjie Yang, Xiaojie Jin, Jiashi Feng, Xiaojun Chang, and Heng Wang. Video recognition in portrait mode. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 21831–21841, 2024. 5

[50] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 770–778, 2016. 1

[51] Jie Hong, Zeeshan Hayder, Junlin Han, Pengfei Fang, Mehrtash Harandi, and Lars Petersson. Hyperbolic audio-visual zero-shot learning. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 7873–7883, 2023. 2

[52] Qiwei Hou, Hongzhi Wu, Zilei Ye, Qian Qiu, Xiaokang Yang, and Meng Wang Tang. Vegfru: A domain-specific dataset for fine-grained visual categorization. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 541–549, 2017. 5

[53] Thomas Hoyoux, Antonio J Rodríguez-Sánchez, and Justus H Piater. Can computer vision problems benefit from structured hierarchical classification? *Machine Vision and Applications*, 27(8):1299–1312, 2016. 1, 2

[54] Sarah Ibrahimi, Mina Ghadimi Atigh, Nanne Van Noord, Pascal Mettes, and Marcel Worring. Intriguing properties of hyperbolic embeddings in vision-language models. *Transactions on Machine Learning Research*, 2024. 2

[55] Jeremy Irvin, Pranav Rajpurkar, Michael Ko, Yifan Yu, Silviana Ciurea-Ilcus, Christopher Chute, Henrik Marklund, Behzad Haghgoo, Robyn Ball, Katie Shpanskaya, Jake Seekins, David Mong, Safwan Halabi, Jacob Sandberg, Russell Jones, David Larson, Curtis Langlotz, B Patel, and Matthew Lungren. Chexpert: A large chest radiograph dataset with uncertainty labels and expert comparison. *Radiology: Artificial Intelligence*, 1(1):e180057, 2019. 5

[56] Menglin Jia, Mengyun Shi, Mikhail Sirotenko, Yin Cui, Claire Cardie, Bharath Hariharan, Hartwig Adam, and Serge Belongie. Fashionpedia: Ontology, segmentation, and an attribute localization dataset. In *Computer vision–ECCV 2020: 16th European conference, glasgow, UK, August 23–28, 2020, proceedings, part i 16*, pages 316–332. Springer, 2020. 5

[57] Tejaswi Kasarla, Max van Spengler, and Pascal Mettes. Balanced hyperbolic embeddings are natural out-of-distribution detectors. *arXiv preprint arXiv:2506.10146*, 2025. 1, 2, 7

[58] Will Kay, Joao Carreira, Karen Simonyan, Brian Zhang, Chloe Hillier, Sudheendra Vijayanarasimhan, Fabio Viola, Tim Green, Trevor Back, Paul Natsev, Mustafa Suleyman, and Andrew Zisserman. The kinetics human action video dataset. *arXiv preprint arXiv:1705.06950*, 2017. 1, 5

[59] Jonathan Krause, Michael Stark, Jia Deng, and Li Fei-Fei. 3d object representations for fine-grained categorization. In *4th International IEEE Workshop on 3D Representation and Recognition (3dRR-13)*, Sydney, Australia, 2013. 5

[60] Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, Li-Jia Li, David Shamma, Michael Bernstein, and Li Fei-Fei. Visual genome: Connecting language and vision using crowdsourced dense image annotations. *International Journal of Computer Vision*, 123(1):32–73, 2017. 4, 5

[61] Alex Krizhevsky and Geoffrey Hinton. Learning multiple layers of features from tiny images. Technical report, University of Toronto, 2009. 2, 5, 7

[62] Alina Kuznetsova, Hassan Rom, Neil Alldrin, Jasper Uijlings, Ivan Krasin, Jordi Pont-Tuset, Shahab Kamali, Stefan Popov, Marco Malloci, Alexander Kolesnikov, Tom Duerig, 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, 5

[63] Marc Law, Renjie Liao, Jake Snell, and Richard Zemel. Lorentzian distance learning for hyperbolic representations. In *International Conference on Machine Learning*, pages 3672–3681. PMLR, 2019. 2

[64] Diego Lazcano, Nicolás Fredes Franco, and Werner Creixell. Hgan: Hyperbolic generative adversarial network. *IEEE Access*, 9:96309–96320, 2021. 2

[65] Fei-Fei Li, Marco Andreeto, Marc'Aurelio Ranzato, and Pietro Perona. Caltech 101, 2022. 5

[66] Li-Jia Li, Chong Wang, Yongwhan Lim, David M Blei, and Li Fei-Fei. Building and using a semantivisual image hierarchy. In *2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pages 3336–3343. IEEE, 2010. 2

[67] Randolph Linderman, Jingyang Zhang, Nathan Inkawhich, Hai Li, and Yiran Chen. Fine-grain inference on out-of-distribution data with hierarchical classification. In *Conference on Lifelong Learning Agents*, pages 162–183. PMLR, 2023. 5

[68] Lu Ling, Yichen Sheng, Zhi Tu, Wentian Zhao, Cheng Xin, Kun Wan, Lantao Yu, Qianyu Guo, Zixun Yu, Yawen Lu, et al. Dl3dv-10k: A large-scale scene dataset for deep learning-based 3d vision. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 22160–22169, 2024. 5

[69] 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 *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9273–9281, 2020. 2

[70] Yang Liu, Lei Zhou, Pengcheng Zhang, Xiao Bai, Lin Gu, Xiaohan Yu, Jun Zhou, and Edwin R Hancock. Where to focus: Investigating hierarchical attention relationship for fine-grained visual classification. In *European Conference on Computer Vision*, pages 57–73. Springer, 2022. 6

[71] Vincenzo Lomonaco and Davide Maltoni. Core50: a new dataset and benchmark for continuous object recognition. In *Conference on Robot Learning*, pages 17–26, 2017. 5

[72] Teng Long, Pascal Mettes, Heng Tao Shen, and Cees GM Snoek. Searching for actions on the hyperbole. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 1141–1150, 2020. 2

[73] 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. 5

[74] Zhitao Long, Deng Cai, and Wenqian Gan. Searching for an effective and efficient hyperbolic representation for supervised learning. In *Advances in Neural Information Processing Systems (NeurIPS)*, pages 18776–18788, 2020. 1, 2, 7

[75] Chao Ma, Jia-Bin Huang, Xiaokang Yang, and Ming-Hsuan Yang. Hierarchical convolutional features for visual tracking. In *Proceedings of the IEEE international conference on computer vision*, pages 3074–3082, 2015. 2

[76] Farzaneh Mahdisoltani, Guillaume Berger, Waseem Gharbieh, David Fleet, and Roland Memisevic. On the effectiveness of task granularity for transfer learning. *arXiv preprint arXiv:1804.09235*, 2018. 5

[77] Subhransu Maji, Raviv Raich, Greg Shakhnarovich, Rogerio Feris, Ludwig Schmidt, Ivan Laptev, and Josef Sivic. Fine-grained visual classification of aircraft. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, 2013. 5

[78] Marcin Marszalek and Cordelia Schmid. Semantic hierarchies for visual object recognition. In *2007 IEEE Conference on Computer Vision and Pattern Recognition*, pages 1–7. IEEE, 2007. 2

[79] Patricia Martinez-Gonzalez, Hedvig Kjellstrom, and Javier J Romero. A grocery store image dataset with visual and semantic labels for object recognition. In *2019 IEEE Winter Conference on Applications of Computer Vision (WACV)*, 2019. 5

[80] 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. 1, 2

[81] Antoine Miech, Dimitri Zhukov, Jean-Baptiste Alayrac, Makarand Tapaswi, Ivan Laptev, and Josef Sivic. HowTo100M: Learning a Text-Video Embedding by Watching Hundred Million Narrated Video Clips. In *ICCV*, 2019. 5

[82] George A Miller. Wordnet: a lexical database for english. *Communications of the ACM*, 38(11):39–41, 1995. 1, 2

[83] Safaa Abdullahi Moallim Mohamud, Minjin Baek, and Dong Seog Han. Hierarchical question-answering for driving scene understanding using vision-language models. *arXiv preprint arXiv:2506.02615*, 2025. 2

[84] Mathew Monfort, Alex Andonian, Bolei Zhou, Kandan Ramakrishnan, Sarah Bargal, Tian Yan, Yinan Lin, Lisa Brown, Qingqiu Fan, Daniel Gutfruend, Carl Vondrick, Aude Oliva, and Antonio Torralba. Moments in time dataset: one million videos for event understanding. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 739–748, 2019. 5

[85] F Morin and Y Bengio. Hierarchical probabilistic neural network language model. In *Proceedings of the Tenth International Workshop on Artificial Intelligence and Statistics (AISTATS)*, pages 246–252, 2005. 2, 6

[86] M. Müller, T. Röder, M. Clausen, B. Eberhardt, B. Krüger, and A. Weber. Documentation mocap database hdm05. Technical Report CG-2007-2, Universität Bonn, 2007. 5

[87] Gerhard Neuhold, Tobias Ollmann, Samuel Rota Bulo, and Peter Kontschieder. The mapillary vistas dataset for semantic urban scene understanding. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 4990–4999, 2017. 5

[88] Maximilian Nickel and Douwe Kiela. Poincaré embeddings for learning hierarchical representations. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2017. 1, 2, 7

[89] 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

[90] Zachary Novack, Julian McAuley, Zachary Chase Lipton, and Saurabh Garg. Chils: Zero-shot image classification with hierarchical label sets. In *International Conference on Machine Learning*, pages 26342–26362. PMLR, 2023. 2

[91] Hyun Oh Song, Yu Xiang, Stefanie Jegelka, and Silvio Savarese. Deep metric learning via lifted structured feature embedding. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4004–4012, 2016. 5

[92] 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. In *The Thirteenth International Conference on Learning Representations*, 2025. 2

[93] Seulki Park, Youren Zhang, X Yu Stella, Sara Beery, and Jonathan Huang. Visually consistent hierarchical image classification. In *The Thirteenth International Conference on Learning Representations*, 2025. 1, 2, 6

[94] 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*, pages 4222–4232, 2025. 2

[95] Eric Qu and Dongmian Zou. Autoencoding hyperbolic representation for adversarial generation. *arXiv preprint arXiv:2201.12825*, 2022. 2

[96] Vignesh Ramanathan, Anmol Kalia, Vladan Petrovic, Yi Wen, Baixue Zheng, Baishan Guo, Rui Wang, Aaron Marquez, Rama Kovvuri, Abhishek Kadian, et al. Paco: Parts and attributes of common objects. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7141–7151, 2023. 5

[97] Sarah Rastegar, Yuki M Asano, Hazel Doughty, and Cees G. M. Snoek. Generalized category discovery with hierarchical label smoothing, 2024. 7

[98] Tal Ridnik, Ethan Ben-Baruch, Assaf Noy, and Lihi Zelnik-Manor. Imagenet-21k pretraining for the masses. *arXiv preprint arXiv:2104.10972*, 2021. 4, 5

[99] Luuk Romeijn, Andrius Bernatavicius, and Duong Vu. Mycoai: Fast and accurate taxonomic classification for fungal its sequences. *Molecular Ecology Resources*, 24(8): e14006, 2024. 7

[100] Amruth Sagar, Rishabh Srivastava, Venkata Kesav Venna, Ravi Kiran Sarvadevabhatla, et al. Madverse: A hierarchical dataset of multi-lingual ads from diverse sources and categories. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 8087–8096, 2024. 5

[101] Frederic Sala, Chris De Sa, Albert Gu, and Christopher Ré. Representation tradeoffs for hyperbolic embeddings. In *International conference on machine learning*, pages 4460–4469. PMLR, 2018. 1, 2, 7

[102] Shibani Santurkar, Dimitris Tsipras, and Aleksander Madry. Breeds: Benchmarks for subpopulation shift. In *International Conference on Learning Representations*, 2021. 2

[103] Rik Sarkar. Low distortion delaunay embedding of trees in hyperbolic plane. In *International symposium on graph drawing*, pages 355–366. Springer, 2011. 1, 7

[104] Shuai Shao, Zeming Zhao, Bo Li, Tiancheng Xiao, Gang Yu, Xiangyu Zhang, and Jian Sun. Objects365: A large-scale, high-quality dataset for object detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 8430–8439, 2019. 4, 5

[105] Qi She, Liang Zhang, Yongchao Zhuang, Haotian Zhao, Jun Yang, Guohao Ma, Jiawei Liang, Jiaqi Dong, Mingcheng Cao, Bin Han, Minghao Wang, Chengdong Jiang, Yi Zhang, Yi Hu, Yongxiang Liu, and Jiaqi Li. Openloris-object: A robotic vision dataset and benchmark for lifelong learning. *International Journal of Computer Vision*, 128: 2413–2430, 2020. 5

[106] Carlos N Silla and Alex A Freitas. A survey of hierarchical classification across different application domains. *Data mining and knowledge discovery*, 22:31–72, 2011. 6

[107] Josef Sivic, Bryan C Russell, Andrew Zisserman, William T Freeman, and Alexei A Efros. Unsupervised discovery of visual object class hierarchies. In *2008 IEEE Conference on Computer Vision and Pattern Recognition*, pages 1–8. IEEE, 2008. 2

[108] Khurram Soomro, Amir Roshan Zamir, and Mubarak Shah. Ucf101: A dataset of 101 human actions classes from videos in the wild. *arXiv preprint arXiv:1212.0402*, 2012. 5

[109] Satwik Srivastava and Deepak Mishra. Severity of error in hierarchical datasets. *Scientific Reports*, 13(1), 2023. 2

[110] Olivia Stevens, David Gunning, Wenjie Miao, Wenbin Zeng, Ana Saldanha, Leslie Smith, Lavanya Baskaran, Pramod Mathai, Deep Joshi, Wei Peng, et al. Bioclip: A vision foundation model for the tree of life. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024. 2, 4, 5

[111] Tanuj Sur, Samrat Mukherjee, Kaizer Rahaman, Subhasis Chaudhuri, Muhammad Haris Khan, and Biplab Banerjee. Hyperbolic uncertainty-aware few-shot incremental point cloud segmentation. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 11810–11821, 2025. 2

[112] Abraham Ungar. *A gyrovector space approach to hyperbolic geometry*. Springer Nature, 2022. 2

[113] Grant Van Horn, Steve Branson, Ryan Farrell, Scott Haber, Jessie Barry, Panos Ipeirotis, Pietro Perona, and Serge Belongie. Building a bird recognition app and large scale dataset with citizen scientists: The fine print in fine-grained dataset collection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 595–604, 2015. 2, 5

[114] Grant Van Horn, Oisin Mac Aodha, Yang Song, Chenyi Cui, Chen Sun, Andrew Shepard, Hartwig Adam, Pietro Perona, and Serge Belongie. The inaturalist species classification and detection dataset. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 8769–8778, 2018. 2, 5

[115] Max van Spengler and Pascal Mettes. Low-distortion and gpu-compatible tree embeddings in hyperbolic space. *arXiv preprint arXiv:2502.17130*, 2025. 1, 2

[116] Max Van Spengler, Erwin Berkhout, and Pascal Mettes. Poincaré resnet. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 5419–5428, 2023. 2

[117] Catherine Wah, Steve Branson, Peter Welinder, Pietro Perona, and Serge Belongie. The caltech-ucsd birds-200-2011 dataset. Technical report, California Institute of Technology, 2011. 1, 2, 5

[118] Jonatas Wehrmann, Ricardo Cerri, and Rodrigo Barros. Hierarchical multi-label classification networks. In *International conference on machine learning*, pages 5075–5084. PMLR, 2018. 6

[119] Xinpeng Wu, Peng Chen, Shuhan Wang, Bing Xu, Qiang Zhang, Yanan Zheng, Yihua Zhao, Yuliang Zhang, Jiawei Ren, Zhiyong Liu, and Guijun Yang. Ip102: A large-scale benchmark dataset for insect pest recognition. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 9177–9184, 2019. 5

[120] Yongqin Xian, Christoph H. Lampert, Bernt Schiele, and Zeynep Akata. Zero-shot learning—a comprehensive evaluation of the good, the bad and the ugly. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 41(9):2251–2265, 2019. 5

[121] Jianxiong Xiao, James Hays, Krista A Ehinger, Aude Oliva, and Antonio Torralba. Sun database: Large-scale scene recognition from abbey to zoo. In *2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pages 3485–3492, 2010. 5

[122] Jianxiong Xiao, Krista A Ehinger, James Hays, Antonio Torralba, and Aude Oliva. Recognizing scene viewpoint using panoramic place representation. In *2012 IEEE Conference on Computer Vision and Pattern Recognition*, pages 2695–2702, 2012. 5

[123] Saining Xie, Chen Sun, Jonathan Huang, Zhuowen Tu, and Kevin Murphy. Rethinking spatiotemporal feature learning: Speed-accuracy trade-offs in video classification. In *Proceedings of the European conference on computer vision (ECCV)*, pages 305–321, 2018. 5

[124] Jinglin Xu, Guohao Zhao, Sibo Yin, Wenhao Zhou, and Yuxin Peng. Finesports: A multi-person hierarchical sports video dataset for fine-grained action understanding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 21773–21782, 2024. 5

[125] Ke Yan, Xiaoshuang Wang, Le Lu, and Ronald M Summers. Deeplesion: automated mining of large-scale lesion annotations and universal lesion detection with deep learning. *Journal of Medical Imaging*, 5(3):036501, 2018. 5

[126] Zhicheng Yan, Hao Zhang, Robinson Piramuthu, Vignesh Jagadeesh, Dennis DeCoste, Wei Di, and Yizhou Yu. Hd-cnn: hierarchical deep convolutional neural networks for large scale visual recognition. In *Proceedings of the IEEE international conference on computer vision*, pages 2740–2748, 2015. 1, 6

[127] Chih-Hsuan Yang, Benjamin Feuer, Talukder Jubery, Zi Deng, Andre Nakkab, Md Zahid Hasan, Shivani Chiranjeevi, Kelly Marshall, Nirmal Baishnab, Asheesh Singh, et al. Biotrove: A large curated image dataset enabling ai for biodiversity. *Advances in Neural Information Processing Systems*, 37:102101–102120, 2024. 5

[128] William Yang, Byron Zhang, and Olga Russakovsky. Imagenet-ood: Deciphering modern out-of-distribution detection algorithms. In *The Twelfth International Conference on Learning Representations*, 2024. 5

[129] 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. 2

[130] Yun Yue, Fangzhou Lin, Kazunori D Yamada, and Ziming Zhang. Hyperbolic contrastive learning. *arXiv preprint arXiv:2302.01409*, 2023. 2

[131] Hanlei Zhang, Hua Xu, Xin Wang, Qianrui Zhou, Shaojie Zhao, and Jiayan Teng. Mintrec: A new dataset for multimodal intent recognition. In *Proceedings of the 30th ACM International Conference on Multimedia*, page 1688–1697, 2022. 5

[132] Yuanhan Zhang, Qinghong Sun, Yichun Zhou, Zexin He, Zhenfei Yin, Kun Wang, Lu Sheng, Yu Qiao, Jing Shao, and Ziwei Liu. Bamboo: Building mega-scale vision dataset continually with human-machine synergy, 2022. 4, 5

[133] Bolei Zhou, Hang Zhao, Xavier Puig, Sanja Fidler, Adela Barriuso, and Antonio Torralba. Scene parsing through ade20k dataset. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 633–641, 2017. 2, 4, 5

[134] Song-Chun Zhu, David Mumford, et al. A stochastic grammar of images. *Foundations and Trends® in Computer Graphics and Vision*, 2(4):259–362, 2007. 2