Improving Visual Recognition with Hyperbolical Visual Hierarchy Mapping

Hyeongjun Kwon[1] Jinhyun Jang[1] Jin Kim[1] Kwonyoung Kim[1] Kwanghoon Sohn[1,2*]
[1]Yonsei University, [2]Korea Institute of Science and Technology (KIST)
{kwonjunn01, jr000192, kimjin928, kyk12, khsohn}@yonsei.ac.kr,

Abstract

Visual scenes are naturally organized in a hierarchy, where a coarse semantic is recursively comprised of several fine details. Exploring such a visual hierarchy is crucial to recognize the complex relations of visual elements, leading to a comprehensive scene understanding. In this paper, we propose a Visual Hierarchy Mapper (Hi-Mapper), a novel approach for enhancing the structured understanding of the pre-trained Deep Neural Networks (DNNs). Hi-Mapper investigates the hierarchical organization of the visual scene by 1) pre-defining a hierarchy tree through the encapsulation of probability densities; and 2) learning the hierarchical relations in hyperbolic space with a novel hierarchical contrastive loss. The pre-defined hierarchy tree recursively interacts with the visual features of the pre-trained DNNs through hierarchy decomposition and encoding procedures, thereby effectively identifying the visual hierarchy and enhancing the recognition of an entire scene. Extensive experiments demonstrate that Hi-Mapper significantly enhances the representation capability of DNNs, leading to an improved performance on various tasks, including image classification and dense prediction tasks. The code is available at https://github.com/kwonjunn01/Hi-Mapper.

1. Introduction

Recognizing and representing the visual scene of any content is the fundamental pursuit of the computer vision field [1–4]. In particular, understanding *what constitutes a scene* and *how each element is comprised of* plays a key role in various visual recognition tasks such as image retrieval [1, 2], human-object interaction [3, 5], and dense prediction [6, 7]. This goes beyond merely learning discriminative feature representations, as it requires to reason about the fine details as well as their associations to comprehend the structured nature of the complex visual scene.

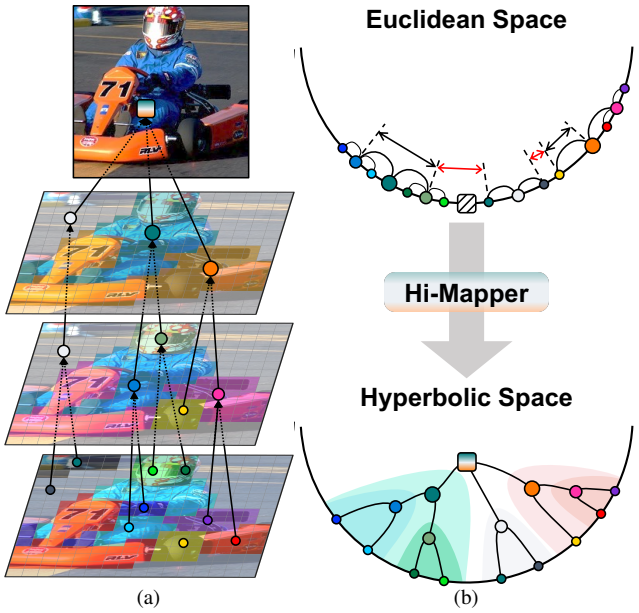


Figure 1. (a) A visual scene can be decomposed into a hierarchical structure based on the semantics of each visual element. (b) Euclidean space is suboptimal in representing the hierarchical structure due to its flat nature. The relational distance is inaccurately captured, being unaware of the semantic similarity of visual elements (Red line). Hi-Mapper maps the hierarchical elements in hyperbolic space, which effectively preserves their semantic relations and distances due to its constant negative curvature.

Over the decades, the development of deep neural networks (DNNs) has contributed towards advances in representing the complex visual scene. Notably, convolutional neural networks (CNNs) have achieved capturing fine details through the local convolutional filters while Vision Transformer (ViT) [8] has enabled coarse context modeling with multi-head self-attention mechanisms. Owing to their different desirable properties, hybrid architectures [9–12] and multi-scale variants of ViTs [13–15] have been extensively explored to capitalize on the complementary features of CNNs and ViTs. Subsequent works [16, 17] have further imposed interaction between multi-scale image patches to facilitate information exchange between fine details and coarse semantics.

*Corresponding author.

This research was supported by the National Research Foundation of Korea (NRF) grant funded by the Korea government (MSIP) (NRF2021R1A2C2006703).

While they have shown effective in capturing coarse-to-fine information, a structured understanding of the visual scene remains underexplored. Concretely, a scene can be interpreted as a hierarchical composition of visual elements, where the ability to recognize an instance (*e.g*. a man) at a coarse level arises from the ability to compose its constituents (*e.g*. body parts) at a finer level, as shown in Fig. 1a. Being aware of such semantic hierarchies furthers the perception of grid-like geometry and enhances the recognition of an entire image. Drawn from the same motivation, the recent variants of ViT [18–20] have constructed the hierarchies between image tokens across the transformer layers. However, their hierarchical relations are defined through the symmetric measurement (*i.e*., cosine similarity) which lacks the ability to represent the asymmetric property of hierarchical structure (*i.e*., inclusion of parent-child nodes). In addition, their image tokens are represented in Euclidean space where the hierarchical relations become distorted due to its linear and flat geometry [21–29], as shown in Fig. 1b.

In this paper, we propose a novel Visual Hierarchy Mapper (Hi-Mapper) that improves the structured understanding of pre-trained DNNs by identifying the visual hierarchy. Hi-Mapper accomplishes this through: 1) Probabilistic modeling of hierarchy nodes, where the mean vector and covariance represent the center and scale of visual-semantic cluster, respectively [30–33]. Accordingly, the asymmetric hierarchical relations are captured through the inclusion of probability densities. Furthermore, 2) Hi-Mapper maps the hierarchy nodes to hyperbolic space, where its constant negative curvature effectively represents the exponential growth of hierarchy nodes. Specifically, Hi-Mapper pre-defines a tree-like structure, with its leaf-level node modeled as a unique Gaussian distribution and the higher-level nodes approximated by a Mixture of Gaussians (MoG) of their corresponding child nodes. The pre-defined hierarchy nodes then interact with the penultimate visual feature map of the pre-trained DNNs to decompose the feature map into the visual hierarchy. Moreover, in order to bypass the difficulties of modeling hierarchy in Euclidean space, Hi-Mapper maps the identified visual hierarchy to hyperbolic space and learns the hierarchical relation with a novel hierarchical contrastive loss. The proposed loss enforces the child-parent nodes to be similar and child-child nodes to be dissimilar in a shared hyperbolic space. The visual hierarchy then interacts with the global visual feature of the pre-trained DNNs such that the hierarchical relations are fully encoded in the global feature representation.

Hi-Mapper serves as a plug-and-play module, which generalizes over any type of DNNs and flexibly identifies the hierarchical organization of visual scenes. We conduct extensive experiments with various pre-trained DNNs (*i.e*., ResNet [34], DeiT [35]) on several benchmarks [36–38] to demonstrate the effectiveness of Hi-Mapper.

In summary, our key contributions are as follows:

- We present a novel Visual Hierarchy Mapper (Hi-Mapper) that enhances the structured understanding of the pre-trained DNNs by investigating the hierarchical organization of visual scene. The proposed Hi-Mapper is applicable to any type of the pre-trained DNNs without modifying the underlying structures.
- Hi-Mapper effectively identifies the visual hierarchy by combining the favorable characteristic of probabilistic modeling and hyperbolic geometry for representing the hierarchical structure.
- We conduct extensive experiments to validate the efficacy of the proposed Hi-Mapper, and improves over the state-of-the-art approaches on various visual recognition tasks.

2. Related Work

Hierarchy-aware visual recognition. Unsupervised image parsing is a long-standing pursuit in visual recognition tasks from the classical computer vision era [39–43]. In the pre-deep learning era, Zhouwen *et al*. [42] firstly introduces a framework to parse images with their constituents via a divide-and-conquer strategy. CapsuleNet-based methods have demonstrated substantial enhancements in image parsing, facilitated by dynamic routing, which efficiently capture the compositional relationships among the activities of capsules that represent object parts. Recently, hierarchical semantic segmentation has been extensively researched, including human parser [44, 45] based on human-part hierarchy and unsupervised part segmentations [46–48].

Beyond image parsing, recent researches on deep neural networks (DNNs) have attempted to exploit hierarchical relationships between detail and global representations. CrossViT [16] utilizes a dual-branch transformer for multi-scale feature extraction, enriching features through a fusion module that integrates inter-scale patch relationships. Quadtree [18] iteratively and hierarchically selects a subset of crucial finer patches within each coarse patch. DependencyViT [19] inverts the self-attention process to organize patches as parent and child nodes, enabling a hierarchical exploration. More recently, CAST [49] employs superpixel-based patch generation and graph pooling for hierarchical patch merging for improving fine-grained recognition performances. While they define hierarchical relations with token similarities in Euclidean space, we pre-define a hierarchical structure with probabilistic modeling and learn the relation in hyperbolic space.

Probabilistic modeling. Probabilistic representation has been extensively explored in the early NLP studies for handling the nuance of word semantics with the probability distribution. For instance, Vilnis *et al*. [30] first introduced the probability densities for representing

word embeddings. Athiwaratkun *et al*. [31] discovered that an imbalance in word frequency leads to distortions in word order and mitigated the problem by representing the word orders through the encapsulation of probability densities. Besides word representation, abundant research has demonstrated the effectiveness of probabilistic modeling in visual representation [6, 50–52]. For example, Shi *et al*. [51] proposed to penalize the low quality face images by measuring the variance of each image distribution. Chun *et al*. [50] identified the limitations of deterministic modeling in vision-language domains and introduced probabilistic cross-modal embedding for providing the uncertainty estimates.

In this work, we deploy probabilistic modeling in defining hierarchical structure, where each distribution represents the inclusive relations of hierarchy nodes.

Hyperbolic manifold. Hyperbolic manifolds have gained increasing interest in deep learning area due to their effectiveness in modeling hierarchical structures. Their success in NLP field [24, 26, 27, 53] has inspired approaches to adopt hyperbolic manifolds in computer vision researches such as image retrieval [1, 2], image segmentation [54, 55], and few-shot learning [25]. As a pioneering work, Khrulkov *et al*. [56] investigated an exponential map from Euclidean space to hyperbolic space for learning hierarchical image embeddings. Ermolov *et al*. [1] applied pair-wise cross entropy loss in hyperbolic space for ViTs. Kim *et al*. [2] extended the work by discovering the latent hierarchy of training data with learnable hierarchical proxies in hyperbolic space. Focusing on pixel-level analysis, [55] identified the long-tail objects by embedding masked instance regions into hyperbolic manifolds. More recently, Desai *et al*. [29] introduced to learn joint image-text embedding space in hyperbolic manifold. While they explore hyperbolic manifold for representing the categorical hierarchies, we identify the hierarchical structure of visual elements without the part-level annotation through a novel hierarchical contrastive loss.

3. Hyperbolic Geometry

Hyperbolic manifold is a smooth *Riemannian manifold* $\mathcal{M}$ with negative curvature c equipped with a Riemannian metric g. The manifold consists of five isometric models and we utilize the Lorentz model for developing Hi-Mapper due to its training stability. A hyperbolic manifold of n-dimensions can be represented as a sub-manifold of the Lorentz model $\mathbb{R}^{n+1}$ as an upper half of a two-sheeted hyperboloid. In the Lorentz space, every point $\mathbf{x} \in \mathbb{R}^{n+1}$ can be denoted as $[\mathbf{x}_{\text{space}}, x_{\text{time}}]$, where $\mathbf{x}_{\text{space}} \in \mathbb{R}^n$ and $x_{\text{time}} \in \mathbb{R}$. Let $\langle \mathbf{x}, \mathbf{y} \rangle$ be the Euclidean inner product and $\langle \mathbf{x}, \mathbf{y} \rangle_{\mathbb{L}}$ denote the *Lorentzian inner product* which is derived by the Riemannian metric of the Lorentz model $g_{\mathbb{L}}$. Given two vectors $\mathbf{x}, \mathbf{y} \in \mathbb{R}^{n+1}$, the Lorentzian inner product is computed as follows:

$$\langle \mathbf{x}, \mathbf{y} \rangle_{\mathbb{L}} = -x_{\text{time}} y_{\text{time}} + \langle \mathbf{x}_{\text{space}}, \mathbf{y}_{\text{space}} \rangle . \tag{1}$$

The n-dimensional Lorentz model $(\mathbb{L}^n, g_{\mathbb{L}})$ is defined by the manifold $\mathbb{L}^n = \left\{ \mathbf{x} \in \mathbb{R}^{n+1} : \langle \mathbf{x}, \mathbf{x} \rangle_{\mathbb{L}} = -1/c, c > 0 \right\}$ and Riemmanian metric of the Lorentz model $g_{\mathbb{L}}$. Thus, all vectors satisfy the following constraints:

$$-x_{\text{time}}^2 + \|\mathbf{x}_{\text{space}}\|^2 = -1/c. \tag{2}$$

A *geodesic* is the shortest path between two vectors on the manifold. The *Lorentzian distance* on $\mathbb{L}$ is then defined as:

$$D_{\mathbb{L}}(\mathbf{x}, \mathbf{y}) = \operatorname{arccosh}(-\langle \mathbf{x}, \mathbf{y} \rangle_{\mathbb{L}}). \tag{3}$$

The *exponential map* is a way to map vectors from tangent space $\mathcal{T}_z \mathbb{L}^n$ onto hyperbolic manifolds $\mathbb{L}^n$, where $\mathcal{T}_z \mathbb{L}^n$ is a Euclidean space of vectors that are orthogonal to some point $\mathbf{z} \in \mathbb{L}^n$. We map the tangent vector $\mathbf{v} \in \mathcal{T}_z \mathbb{L}^n$ from Euclidean space to the Lorentz manifolds, in which the exponential map $\exp\mathbf{m}_{\mathbf{z}}$ is defined as:

$$\mathbf{x} = \exp\mathbf{m}_z(\mathbf{v}) = \cosh(\sqrt{c}\|\mathbf{v}\|_{\mathbb{L}})\mathbf{z} + \frac{\sinh(\sqrt{c}\|\mathbf{v}\|_{\mathbb{L}})}{\sqrt{c}\|\mathbf{v}\|_{\mathbb{L}}}\mathbf{v}. \tag{4}$$

The *logarithm map* $\log\mathbf{m}_z$ which transfers $\mathbf{x}$ on the hyperboloid back to the tangent space $\mathcal{T}_z \mathcal{M}$ is defined as:

$$\mathbf{v} = \log\mathbf{m}_z(\mathbf{x}) = \frac{\cosh^{-1}(-c\langle \mathbf{z}, \mathbf{x} \rangle_{\mathbb{L}})}{\sqrt{(c\langle \mathbf{z}, \mathbf{x} \rangle_{\mathbb{L}})^2 - 1}} \text{proj}_{\mathbf{z}}(\mathbf{x}). \tag{5}$$

We set $\mathbf{z}$ as the origin of the hyperboloid $\mathbf{O} = \left[\mathbf{0}, \sqrt{1/c}\right]$.

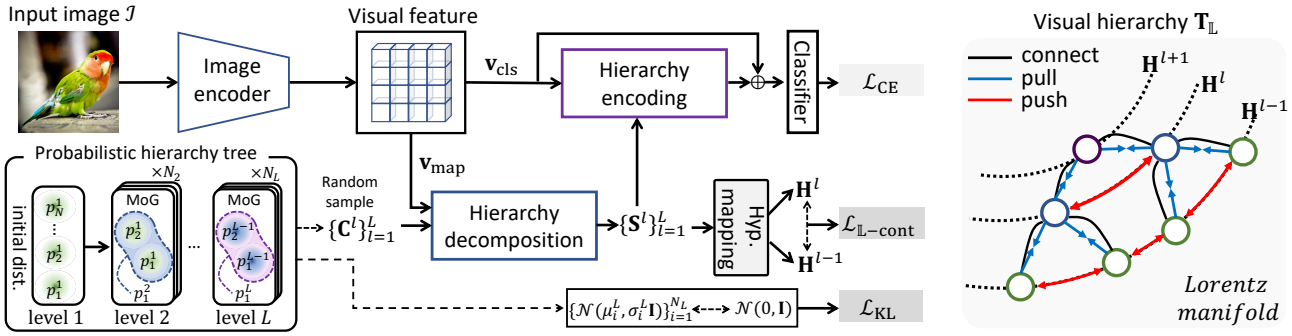
4. Method

4.1. Overview

Our goal is to enhance the structured understanding of pre-trained deep neural networks (DNNs) by investigating the hierarchical organization of visual scenes. To this end, we introduce a Visual Hierarchy Mapper (Hi-Mapper) which serves as a plug-and-play module on any type of pre-trained DNNs. An overview of Hi-Mapper is depicted in Fig. 2a.

Given an image $\mathcal{I}$, we first extract visual features $[\mathbf{v}_{\text{map}}, \mathbf{v}_{\text{cls}}] = \mathcal{F}(\mathcal{I})$ from a pre-trained image encoder $\mathcal{F}$, where $\mathbf{v}_{\text{map}} \in \mathbb{R}^{hw \times d}$ is the penultimate visual feature map and $\mathbf{v}_{\text{cls}} \in \mathbb{R}^d$ is the global visual representation, with hw indicating the size of the visual feature map.

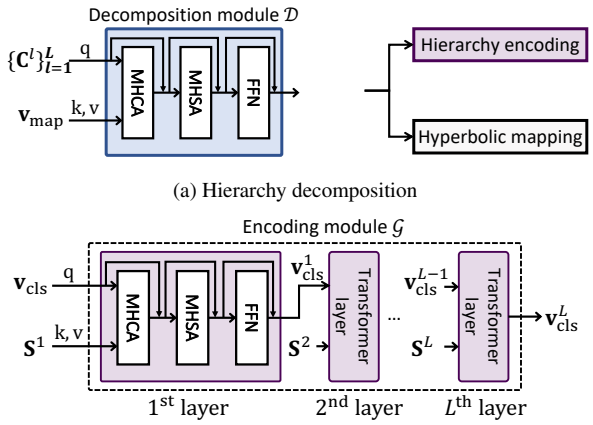
Hi-Mapper identifies the visual hierarchy from the visual feature map $\mathbf{v}_{\text{map}}$ and encodes the identified visual hierarchy back to the global visual representation $\mathbf{v}_{\text{cls}}$ for enhancing the recognition of a whole scene. To this end, we pre-define a hierarchy tree with Gaussian distribution, where the relations of the hierarchy nodes are defined through the inclusion of probability densities (Sec. 4.2). The pre-defined



(a) Overall pipeline

(b) Hierarchical contrastive loss

Figure 2. (a) An overview of the proposed Hi-Mapper. Hi-Mapper operates on top of pre-trained image encoder $\mathcal{F}$, with probabilistic hierarchy tree $\mathbf{T} = \{\mathbf{C}^l\}_{l=1}^L$. The tree interacts with visual feature map $\mathbf{v}_{\text{map}}$ through hierarchy decomposition module $\mathcal{D}$, thereby identifying visual hierarchy in Euclidean space $\mathbf{T}_{\mathbb{E}} = \{\mathbf{S}^l\}_{l=1}^L$. The visual hierarchy is mapped to hyperbolic space $\mathbf{T}_{\mathbb{L}} = \{\mathbf{H}^l\}_{l=1}^L$ and optimized with hierarchical contrastive loss $\mathcal{L}_{\mathbb{L}\text{-cont}}$. The visual hierarchy is further encoded into global visual representation $\mathbf{v}_{\text{cls}}$ via hierarchy encoding module $\mathcal{G}$ for enhancing the recognition of entire scene. (b) The proposed hierarchical contrastive loss pulls each parent-child node and pushes all the other nodes at the same level.



(a) Hierarchy decomposition

(b) Hierarchy encoding

Figure 3. (a) Hierarchy decomposition module groups semantically-relevant visual features $\mathbf{v}_{\text{map}}$ to the closest semantic cluster $\mathbf{C}^l$. (b) Hierarchy encoding module progressively updates global representation $\mathbf{v}_{\text{cls}}$ by aggregating the visual hierarchy $\mathbf{S}^l$.

hierarchy tree interacts with $\mathbf{v}_{\text{map}}$ through the hierarchy decomposition module $\mathcal{D}$ such that the feature map is decomposed into the visual hierarchy (Sec. 4.3). Since the zero curvature of Euclidean space is not optimal for representing the hierarchical structure, we map the visual hierarchy to hyperbolic space and optimize the relation with a novel hierarchical contrastive loss (Sec. 4.4). The visual hierarchy is then encoded back to the global visual representation $\mathbf{v}_{\text{cls}}$ through the hierarchy encoding module $\mathcal{G}$ resulting in an enhanced global representation (Sec. 4.5).

4.2. Probabilistic hierarchy tree

The main problem of the recent hierarchy-aware ViTs [18–20] is that they define the hierarchical relations between the image tokens mainly through the self-attention scores. Such a symmetric measurement is suboptimal for representing the asymmetric inclusive relation of parent-child nodes. To handle the problem, we propose to define L levels of hierarchy tree $\mathbf{T}$ by modeling each hierarchy node with a probability distribution.

Specifically, we first parameterize each leaf-level (at the initial level) node as a unique Gaussian distribution and subsequently define the higher-level node as a Mixture-of-Gaussians (MoG) of its corresponding child nodes. Accordingly, the mean vector represents the cluster center of the visual semantic and the covariance captures the scale of each semantic cluster.

Initial level. Let $\mathbf{C}^1 = \{\mathbf{c}_n^1\}_{n=1}^{N_1}$ be a set of N_1 initial level nodes. We parameterize each node $\mathbf{c}_n^1$ as a normal distribution with a mean vector μ_n^1 and a diagonal covariance matrix $\sigma_n^1\mathbf{I}$ in $\mathbb{R}^d$ as:

$$p(\mathbf{c}_n^1) \sim \mathcal{N}(\mu_n^1, \sigma_n^1\mathbf{I}), \tag{6}$$

where μ_n^1 and σ_n^1 are randomly initialized. We use reparameterization trick [57] for stable sampling, such that:

$$\mathbf{c}_n^1 = \mu_n^1 + \epsilon * \sigma_n^1 \in \mathbb{R}^d, \tag{7}$$

where $\epsilon \sim \mathcal{N}(0, \mathbf{I})$.

Subsequent level. We derive the remaining $(L-1)$ levels of hierarchy tree by conditioning each level on the preceding hierarchy level. For each l-th level, we formulate a set of N_l nodes $\mathbf{C}^l = \{\mathbf{c}_k^l\}_{k=1}^{N_l}$. Concretely, the k-th node $\mathbf{c}_k^l$ at the l-th level is approximated by a MoG of its two corresponding child nodes, $\mathbf{c}_{2k-1}^{l-1}$ and $\mathbf{c}_{2k}^{l-1}$, as:

$$p(\mathbf{c}_k^l) \sim \sum_{k'=2k-1}^{2k} \mathcal{N}(\mu_{k'}^{l-1}, \sigma_{k'}^{l-1}\mathbf{I}). \tag{8}$$

To stabilize the construction of hierarchy tree, we increase the sampling rate as the distribution expands, *i.e.*, we sample 2^{l-1} embeddings from $p(\mathbf{c}_k^l)$ using the reparameterization trick in Eq. (7) as:

$$\begin{aligned} \mathbf{c}_k^l = \{\hat{\mathbf{c}}_{k,1}^l, ..., \hat{\mathbf{c}}_{k,2^{l-1}}^l\} \overset{\text{iid}}{\sim} p(\mathbf{c}_k^l), \quad \mathbf{c}_k^l \in \mathbb{R}^{2^{l-1} \times d}, \\ \mathbf{C}^l = \{\mathbf{c}_1^l, ..., \mathbf{c}_{N_l}^l\} \in \mathbb{R}^{2^{l-1} N_l \times d}. \end{aligned} \tag{9}$$

By sequentially conditioning the higher-level nodes on the preceding lower-level nodes, we obtain the hierarchy tree $\mathbf{T} = \{\mathbf{C}^l\}_{l=1}^L$, where the lower-level nodes capture fine details with concentrated distribution and the higher-level nodes capture coarse instance-level representations with dispersed distribution.

KL divergence loss. To prevent the variances from collapsing to zero, we employ KL regularization term between the distributions of L-th level nodes and the unit Gaussian prior $\mathcal{N}(0, \mathbf{I})$ following [50]:

$$\mathcal{L}_{\text{KL}} = \sum_{i=1}^{N_L} \text{KL}(\mathcal{N}(\mu_i^L, \sigma_i^L) \parallel \mathcal{N}(0, \mathbf{I})). \tag{10}$$

4.3. Visual hierarchy decomposition

Given the pre-defined hierarchy tree $\mathbf{T}$ and the visual feature map $\mathbf{v}_{\text{map}}$, we decompose $\mathbf{v}_{\text{map}}$ into L levels of visual hierarchy through hierarchy decomposition module $\mathcal{D}$, as shown in Fig. 3a. We instantiate $\mathcal{D}$ as a stack of two transformer decoder layers.

To identify the visual hierarchy at the l-th level, the decomposition module $\mathcal{D}$ treats $\mathbf{C}^l$ as the query, and $\mathbf{v}_{\text{map}}$ as the key and value such that the semantically-relevant visual features are aggregated to the closest semantic cluster:

$$\mathbf{S}^l = \{\mathbf{s}_1^l, ..., \mathbf{s}_{N_l}^l\} = \mathcal{D}(\mathbf{C}^l, \mathbf{v}_{\text{map}}) \in \mathbb{R}^{2^{l-1} N_l \times d}. \tag{11}$$

Similar to the hierarchy nodes $\mathbf{c}_k^l$ in Eq. (9), the decomposed visual hierarchy nodes $\mathbf{s}_k^l$ are comprised of 2^{l-1} visual representations as $\mathbf{s}_k^l = \{\hat{\mathbf{s}}_{k,1}^l, ..., \hat{\mathbf{s}}_{k,2^{l-1}}^l\}$. We average the set of 2^{l-1} representations for each visual hierarchy node $\mathbf{s}_k^l$ such that:

$$\mathbf{s}_k^l = \frac{1}{2^{l-1}} \sum_{i=1}^{2^{l-1}} \hat{\mathbf{s}}_{k,i}^l, \quad \mathbf{s}_k^l \in \mathbb{R}^d \tag{12}$$

We perform the same decomposition procedure for L levels, thereby obtaining L levels of visual hierarchy in Euclidean space $\mathbf{T}_{\mathbb{E}} = \{\mathbf{S}^l\}_{l=1}^L$.

4.4. Learning hierarchy in hyperbolic space

A natural characteristic of the hierarchical structure is that the number of nodes exponentially increases as the depth increases. In practice, representing this property in Euclidean space leads to distortions in the semantic distances due to its flat geometry. We propose to handle the problem by learning the hierarchical relations in hyperbolic space, where its exponentially expanding volume can efficiently represent the visual hierarchy.

We first map $\mathbf{T}_{\mathbb{E}}$ to the Lorentz hyperboloid to derive the visual hierarchy in hyperbolic space $\mathbf{T}_{\mathbb{L}}$. Following [29], we simplify the mapping computation by parameterizing only the space component of the Lorentz model as $\mathbf{s} = [\mathbf{s}_k^l, 0] \in \mathbb{R}^{d+1}$, where $\mathbf{s}$ belongs to the tangent space at the hyperboloid origin $\mathbf{O}$. The visual hierarchy node in hyperbolic space $\mathbf{h}_k^l = [\mathbf{h}_{k,\text{space}}^l, h_{k,\text{time}}^l]$ is then obtained by transforming $\mathbf{s}_k^l$ to $\mathbf{h}_{k,\text{space}}^l$ using the exponential map in Eq. (13) as:

$$\begin{aligned} \mathbf{h}_{k,\text{space}}^l &= \cosh(\sqrt{c}\|\mathbf{s}\|_{\mathbb{L}})\mathbf{0} + \frac{\sinh(\sqrt{c}\|\mathbf{s}\|_{\mathbb{L}})}{\sqrt{c}\|\mathbf{s}\|_{\mathbb{L}}} \mathbf{s}_k^l \\ &= \frac{\sinh(\sqrt{c}\|\mathbf{s}_k^l\|)}{\sqrt{c}\|\mathbf{s}_k^l\|} \mathbf{s}_k^l, \end{aligned} \tag{13}$$

and computing the corresponding time component $h_{k,\text{time}}^l$ using Eq. (2) as $h_{k,\text{time}}^l = \sqrt{1/c + \|\mathbf{h}_{k,\text{space}}^l\|^2}$. The visual hierarchy in hyperbolic space $\mathbf{T}_{\mathbb{L}}$ is denoted as:

$$\mathbf{T}_{\mathbb{L}} = \{\mathbf{H}^l\}_{l=1}^L, \quad \mathbf{H}^l = \{\mathbf{h}_1^l, ..., \mathbf{h}_{N_l}^l\} \in \mathbb{R}^{N_l \times d_{\mathbb{L}}}, \tag{14}$$

where $d_{\mathbb{L}}$ is the embedding dimension in Lorentz manifold.

Hierarchical contrastive loss. To guarantee the hierarchy nodes to reflect the hierarchical relations, we optimize the distances of $\mathbf{T}_{\mathbb{L}}$ with a novel hierarchical contrastive loss $\mathcal{L}_{\mathbb{L}\text{-cont}}$. Specifically, we formulate the similarity between the nodes in consideration of the length of the connected branch, *i.e.*, *geodesic* distance, as shown in Fig. 2b.

Consider a node $\mathbf{h}_{2k}^l$ in $\mathbf{T}_{\mathbb{L}}$, where $\mathbf{h}_k^{l+1}$ is its parent node. We encourage $\mathbf{h}_{2k}^l$ to be similar with $\mathbf{h}_k^{l+1}$ since the pair of parent-child nodes lie on the same branch. Meanwhile, we encourage the remaining nodes in the same level, *i.e.*, $\{\mathbf{h}_i^l | \mathbf{h}_i^l \in \mathbf{H}^l, i \neq 2k\}$, to be dissimilar since they all lie on separate branches. The loss also penalizes the close sibling node $\mathbf{h}_{2k-1}^l$ as their geodesic traverses the parent node. Formally, the hierarchical contrastive loss incorporates the geometric interpretation of hierarchical structure into the contrastive loss [58] as:

$$\mathcal{L}_{\mathbb{L}\text{-cont}} = -\log \mathop{\mathbb{E}}_{\mathbf{T}_{\mathbb{L}}} \left[\sum_{i=0}^{1} \frac{\exp(D_{\mathbb{L}}(\mathbf{h}_k^{l+1}, \mathbf{h}_{2k-i}^l))}{\sum_{j \neq 2k-i} \exp(D_{\mathbb{L}}(\mathbf{h}_{2k-i}^l, \mathbf{h}_j^l))} \right], \tag{15}$$

where $D_{\mathbb{L}}$ is the Lorentzian distance defined in Eq. (3).

By regarding every pair of parent-child nodes as positive and every pair of nodes at the same level as negative, we are able to represent the logical structure of visual hierarchy without the part-level annotation. In addition, the

hierarchical relation optimized through $\mathbf{T}_{\mathbb{L}}$ in hyperbolic space is well-preserved in $\mathbf{T}_{\mathbb{E}}$ in Euclidean space thanks to the mapping computation in Eq. (13).

4.5. Visual hierarchy encoding

We propose to enhance the structured understanding of the global visual representation by encoding the identified visual hierarchy into $\mathbf{v}_{\text{cls}}$. While we optimize the hierarchical relations in hyperbolic space due to its ability of handling *relative* distances, its complex computations and expansion property make it less suitable for tasks that require *absolute* criteria. Thus, we exploit the visual hierarchy in Euclidean space $\mathbf{T}_{\mathbb{E}}$ for hierarchy encoding.

Specifically, we progressively encode the hierarchy nodes $\mathbf{S}^l$ into $\mathbf{v}_{\text{cls}}$, from the initial level to the L-th level, via the hierarchy encoding module $\mathcal{G}$ such that $\mathbf{v}_{\text{cls}}$ is updated as a hierarchy-aware global representation $\mathbf{v}_{\text{cls}}^L$. As shown in Fig. 3b, the module $\mathcal{G}$ is instantiated as a stack of L transformer decoder layers.

To encode the l-th level hierarchy, the l-th layer of $\mathcal{G}^l$ takes $\mathbf{v}_{\text{cls}}^{l-1}$ as the query, and $\mathbf{S}^l$ as the key and value:

$$\mathbf{v}_{\text{cls}}^l = \mathcal{G}^l(\mathbf{v}_{\text{cls}}^{l-1}, \mathbf{S}^l) \in \mathbb{R}^d, \quad 0 < l \leq L, \tag{16}$$

where we set $\mathbf{v}_{\text{cls}}^0$ as $\mathbf{v}_{\text{cls}}$. The enhanced global visual representation is then derived as:

$$\hat{\mathbf{v}}_{\text{cls}} = \mathbf{v}_{\text{cls}} + \mathbf{v}_{\text{cls}}^L. \tag{17}$$

For the final prediction, we feed $\hat{\mathbf{v}}_{\text{cls}}$ into the pre-trained classifier of $\mathcal{F}$.

Overall objectives. The overall objective is defined as:

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{CE}} + \alpha\mathcal{L}_{\mathbb{L}\text{-cont}} + \beta\mathcal{L}_{\text{KL}}, \tag{18}$$

where α and β are the balancing parameters.

5. Experiments

In this section, we conduct extensive experiments to show the effectiveness of our proposed Hi-Mapper on image classification (Sec. 5.1), object detection and instance segmentation (Sec. 5.2), and semantic segmentation (Sec. 5.3). We apply our Hi-Mapper on both CNN-based [34, 59] and ViT-based [35, 60] backbone networks and compare the performance with the concurrent hierarchy-aware baselines [16, 19]. Lastly, we provide ablation studies (Sec. 6) to demonstrate the effectiveness of our contributions.

5.1. Image classification

Settings. We apply our Hi-Mapper on the state-of-the-art backbone networks [34, 35, 59, 60] and benchmark on ImageNet-1k [36] dataset. Following [12, 60, 62–64], we use the identical combinations for data augmentation and regularization strategies [35] after excluding repeated augmentation [65]. We train our model with batch size 1024 for 50 epochs using AdamW [66] with weight decay 1e-4. The initial learning rate is set to 1e-4 and a cosine learning rate scheduler is applied following [35].

Table 1. Performance comparisons for image classification on ImageNet-1K [36] dataset.

Type	Model	Params (M)	FLOPs (G)	Top-1 (%)
CNN	ResNet-50 [34]	25.6	3.8	76.2
	EfficientNet-B4* [59]	19.3	4.2	82.9
	Hi-Mapper(RN50)	26.9	4.0	**78.2**
	Hi-Mapper(ENB4*)	20.5	4.4	**84.1**
ViT	DeiT-T [35]	5.7	1.3	72.2
	CrossViT-T [16]	6.9	1.6	73.4
	PVT-T [61]	13.2	1.9	75.1
	DependecnyViT-T [19]	6.2	1.3	**75.4**
	Hi-Mapper(DeiT-T)	6.6	1.5	74.8
	DeiT-S [35]	22.2	4.5	79.8
	CrossViT-S [16]	26.7	5.6	81
	PVT-S [61]	24.5	3.8	79.8
	Swin-T [60]	28.3	4.6	81.2
	DependencyViT-S [19]	24.0	5.0	82.1
	Hi-Mapper(DeiT-S)	23.3	4.8	**82.6**
	Hi-Mapper(Swin-T)	29.5	4.9	**83.4**
	DeiT-B [35]	85.6	17.6	81.8
	Swin-S [60]	50.1	8.7	83.0
	Hi-Mapper(DeiT-B)	87.2	18.1	**83.4**
	Hi-Mapper(Swin-S)	51.8	9.3	**84.1**

Results. Tab. 1 presents classification performance. We report the original performance without the fine-tuning schemes for the plain backbone networks since we observed degradation in performance after fine-tuning the models. Our proposed method consistently achieves better performances than the baseline models with only a slight increase in parameters. Specifically, Hi-Mapper surpasses ResNet-50 [34] and EfficientNet-B4 [59] by margins of 2.0% and 1.2%, respectively. Additionally, it achieves improvement on DeiT's performance by 2.6%/2.8%/1.6%, and Swin [60] by 2.2%/1.1% across various model sizes. The experiments on image classification demonstrate not only the importance of understanding the structured organization of visual scenes, but also the scalability of our method.

5.2. Object detection and Instance segmentation

Settings. We experiment on the COCO [37] dataset. We use the Hi-Mapper backbones derived from Sec. 5.1. Then,

Table 2. Performance comparisons for object detection and instance segmentation on COCO [37] dataset.(*: Results are reproduced based on [10].

Backbone	Params (M)	Flops (G)	Mask R-CNN 1x					
			AP^{box}	AP^{box}_{50}	AP^{box}_{75}	AP^{mask}	AP^{mask}_{50}	AP^{mask}_{75}
PVT-T [61]	32.9	195	36.7	59.2	39.3	35.1	56.7	37.3
DeiT-T* [35]	27.3	244	30.3	46.2	32.1	27.1	44.4	28.3
Hi-Mapper(Deit-T)	29.5	246	37.1	61.7	41.0	35.7	59.1	38.1
PVT-S [61]	44.1	245	40.4	62.9	43.8	37.8	60.1	40.3
DeiT-S* [35]	44.9	276	36.3	54.1	39.0	31.7	51.3	33.2
Swin-T [60]	47.8	267	43.5	66.4	47.3	39.6	63.1	42.3
Hi-Mapper(Deit-S)	47.0	279	42.6	65.8	46.3	38.9	62.7	41.8
Hi-Mapper(Swin-T)	50.4	270	44.0	67.1	47.9	39.9	63.6	42.8

Table 3. Performance comparisons for semantic segmentation on ADE20k [38] dataset. We conduct the single-scale evaluation. FLOPs are measured with 512 $\times$ 2048 input resolution.

Backbone	Method	Params (M)	FLOPs (G)	mIoU
PVT-T [61]	SemanticFPN [69]	17.0	132	35.7
DeiT-T [35]	UperNet [70]	10.7	142	37.8
DependencyViT-T [19]	UperNet [70]	11.1	145	40.3
Hi-Mapper(Deit-T)	UperNet [70]	11.6	144	39.8
PVT-S [61]	SemanticFPN [69]	28.2	712	39.8
DeiT-S [35]	UperNet [70]	41.3	566	43.0
Swin-T [60]	UperNet [70]	60.0	945	44.5
DependencyVIT-S [19]	UperNet [70]	43.1	574	45.7
Hi-Mapper(Deit-S)	UperNet [70]	42.5	570	46.3
Hi-Mapper(Swin-T)	UperNet [70]	62.1	949	46.8

we deploy our pre-trained backbone into Mask R-CNN [67]. All models are trained on COCO 2017, including 118k train images and 5k validation images. We follow the standard learning protocols [67], 1$\times$ schedule with 12 epochs. Note that we reproduce DeiT-T with Mask R-CNN based on [68].

Results. In Table 2, we present a comparison of our results with baseline models, such as DeiT and Swin, on the COCO dataset, which demands a higher capacity for fine representation recognition. Our Hi-Mapper consistently boosts all baseline models capability with only a small increase in parameters. Notably, Hi-Mapper significantly improves DeiT-T and -S by margins of 6.8 and 8.3 on the object detection task. Meanwhile, in the instance segmentation, it also improves DeiT-T and -S by margins of 8.6 and 7.2, respectively. This result shows that the visual hierarchy facilitates complex visual scene recognition.

5.3. Semantic segmentation

Settings. We further experiment on the ADE20K [38] dataset for semantic segmentation. ADE20K contains 20k training images, 20K validation images, and 3K test images, covering a total 150 classes. Following common practice [70], we report the mIoU on the validation set.

Table 4. Performance comparison for classification on ImageNet-1K [36] according to embedding spaces and combinations of learning objectives.

Manifold	$\mathcal{L}_{\mathbb{L}\text{-cont}}$	$\mathcal{L}_{\text{KL}}$	Top-1(%)
Euclidean	✗	✗	79.7
	✗	✓	79.6
	✓	✗	79.2
	✓	✓	79.3
Hyperbolic	✗	✗	79.7
	✗	✓	79.5
	✓	✗	82.0
	✓	✓	82.6

Results. We present the performance comparisons on ADE20K in Tab. 3. The results show that our Hi-Mapper achieves comparable or better performance than the baseline models, including DeiT-T, -S, and Swin-T, requiring a smaller increase in the number of parameters and GFLOPs. Specifically, Hi-Mapper on DeiT-T, -S, and Swin-T achieves a performance improvement of 2.0%, 3.3%, and 2.3%. Additionally, as model sizes increase, Hi-Mapper fully capitalizes fine-grained representations for semantic segmentation with a slight increase in computation.

5.4. Visualization

As shown in Fig. 4, we demonstrate our visual hierarchy on images from ImageNet-1K [36]. This confirms that our approach can successfully uncover the inherent hierarchy among visual components without the need for hierarchy or part-level annotations.

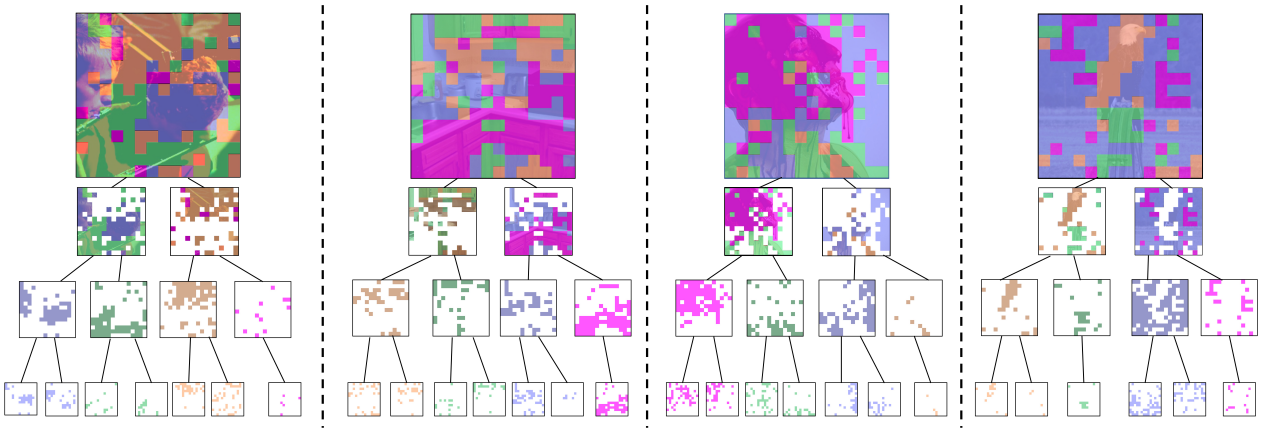


Figure 4. Visualization of visual hierarchy decomposed by Hi-Mapper(Deit-S) trained on ImageNet-1K with classification objective. Each color represents different subtrees. We ignore the nodes of the small region and display only the main subtrees.

Table 5. Performance comparison for classification on ImageNet-1K [36] with respect to the relation modeling in hierarchy tree.

Model	Deterministic	Probabilistic
Hi-Mapper(DeiT-S)	81.5%	82.6%

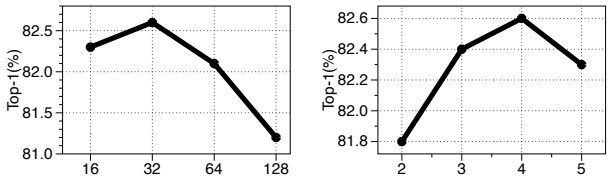
6. Ablation studies and discussion

To further analyze and validate the components of our method, we conduct ablation studies on image classification.

Effectiveness of hyperbolic manifolds. We first investigate the effectiveness of hyperbolic manifolds in our approach. As shown in Tab. 4, we report the impact according to image classification. In Euclidean space, the distance function between two vectors is the cosine similarity function. The results demonstrate that applying hierarchical contrastive loss in Euclidean space degrades performance. It indicates that hyperbolic space is more suitable for stabilizing hierarchical structures. Additionally, the application of a KL loss term shows further benefits derived from the semantic seed distribution.

Impact of probabilistic modeling. In Tab. 5, we report the performance comparisons between the probabilistic hierarchy tree and the deterministic hierarchy tree. For constructing a hierarchy tree, probabilistic modeling defines every node via MoG of its child node distributions, while the deterministic approach determines each node the mean of its child nodes. The probabilistic hierarchy tree achieves significant improvement in performance compared to the deterministic approach. This result shows that probabilistic modeling is more effective in representing hierarchical structure than deterministic modeling, leading to improvement in recognition.

Hierarchy width and depth. As shown in Fig. 5, we analyze the effect of the width N and depth L of the hierarchy tree on ImageNet-1K [36] with Hi-Mapper(DeiT-S). These factors control the granularity of visual elements to be decomposed. While a small number of N degrades the fine-grained recognition capacity, an excess N hinders the optimization. Meanwhile, a large L may provide diverse granularity, however, it leads to entangled object-level representations. In all the cases, we report the best performance of 82.6% at $N = 32, L = 4$.



(a) $N \in \{16, 32, 64, 128\}$ (b) $L \in \{2, 3, 4, 5\}$

Figure 5. Hyper-parameter analysis on ImageNet-1K.

7. Conclusion

In this paper, we have presented a novel Visual Hierarchy Mapper (Hi-Mapper) that investigates the hierarchical organization of visual scenes. We have achieved the goal by newly defining tree-like structure with probability distribution and learning the hierarchical relations in hyperbolic space. We have incorporated the hierarchical interpretation into the contrastive loss and efficiently identified the visual hierarchy in a data-efficient manner. Through an effective hierarchy decomposition and encoding procedures, the identified hierarchy has been successfully deployed to the global visual representation, enhancing the structured understanding of an entire scene. Hi-Mapper has consistently improved the performance of the existing DNNs when integrated with them, and also has demonstrated the effectiveness on various dense predictions.

Acknowledgement. This research was supported by the Yonsei Signature Research Cluster Program of 2022 (2022-22-0002).

References

[1] Aleksandr Ermolov, Leyla Mirvakhabova, Valentin Khrulkov, Nicu Sebe, and Ivan Oseledets. Hyperbolic vision transformers: Combining improvements in metric learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7409–7419, 2022. 1, 3

[2] 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. 1, 3

[3] Georgia Gkioxari, Ross Girshick, Piotr Dollár, and Kaiming He. Detecting and recognizing human-object interactions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 8359–8367, 2018. 1

[4] Jinhyun Jang, Jungin Park, Jin Kim, Hyeongjun Kwon, and Kwanghoon Sohn. Knowing where to focus: Event-aware transformer for video grounding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13846–13856, 2023. 1

[5] Zhi Hou, Baosheng Yu, Yu Qiao, Xiaojiang Peng, and Dacheng Tao. Affordance transfer learning for human-object interaction detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 495–504, 2021. 1

[6] Hyeongjun Kwon, Taeyong Song, Somi Jeong, Jin Kim, Jinhyun Jang, and Kwanghoon Sohn. Probabilistic prompt learning for dense prediction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6768–6777, 2023. 1, 3

[7] Jin Kim, Jiyoung Lee, Jungin Park, Dongbo Min, and Kwanghoon Sohn. Pin the memory: Learning to generalize semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4350–4360, 2022. 1

[8] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020. 1

[9] Prajit Ramachandran, Niki Parmar, Ashish Vaswani, Irwan Bello, Anselm Levskaya, and Jon Shlens. Stand-alone self-attention in vision models. *Advances in neural information processing systems*, 32, 2019. 1

[10] Hengshuang Zhao, Jiaya Jia, and Vladlen Koltun. Exploring self-attention for image recognition. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10076–10085, 2020. 7

[11] Jianyuan Guo, Kai Han, Han Wu, Yehui Tang, Xinghao Chen, Yunhe Wang, and Chang Xu. Cmt: Convolutional neural networks meet vision transformers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12175–12185, 2022.

[12] Haiping Wu, Bin Xiao, Noel Codella, Mengchen Liu, Xiyang Dai, Lu Yuan, and Lei Zhang. Cvt: Introducing convolutions to vision transformers. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 22–31, 2021. 1, 6

[13] Xiaoyi Dong, Jianmin Bao, Dongdong Chen, Weiming Zhang, Nenghai Yu, Lu Yuan, Dong Chen, and Baining Guo. Cswin transformer: A general vision transformer backbone with cross-shaped windows. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12124–12134, 2022. 1

[14] Wenhai Wang, Enze Xie, Xiang Li, Deng-Ping Fan, Kaitao Song, Ding Liang, Tong Lu, Ping Luo, and Ling Shao. Pyramid vision transformer: A versatile backbone for dense prediction without convolutions. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 568–578, 2021.

[15] Yanghao Li, Chao-Yuan Wu, Haoqi Fan, Karttikeya Mangalam, Bo Xiong, Jitendra Malik, and Christoph Feichtenhofer. Mvitv2: Improved multiscale vision transformers for classification and detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4804–4814, 2022. 1

[16] Chun-Fu Richard Chen, Quanfu Fan, and Rameswar Panda. Crossvit: Cross-attention multi-scale vision transformer for image classification. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 357–366, 2021. 1, 2, 6

[17] Pengzhen Ren, Changlin Li, Guangrun Wang, Yun Xiao, Qing Du, Xiaodan Liang, and Xiaojun Chang. Beyond fixation: Dynamic window visual transformer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11987–11997, 2022. 1

[18] Shitao Tang, Jiahui Zhang, Siyu Zhu, and Ping Tan. Quadtree attention for vision transformers. *arXiv preprint arXiv:2201.02767*, 2022. 2, 4

[19] Mingyu Ding, Yikang Shen, Lijie Fan, Zhenfang Chen, Zitian Chen, Ping Luo, Joshua B Tenenbaum, and Chuang Gan. Visual dependency transformers: Dependency tree emerges from reversed attention. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14528–14539, 2023. 2, 6, 7

[20] Tsung-Wei Ke, Sangwoo Mo, and X Yu Stella. Learning hierarchical image segmentation for recognition and by recognition. In *The Twelfth International Conference on Learning Representations*, 2023. 2, 4

[21] N. Linial, E. London, and Y. Rabinovich. The geometry of graphs and some of its algorithmic applications. In *Proceedings 35th Annual Symposium on Foundations of Computer Science*, pages 577–591, 1994. doi: 10.1109/SFCS.1994.365733. 2

[22] Hongbin Pei, Bingzhe Wei, Kevin Chang, Chunxu Zhang, and Bo Yang. Curvature regularization to prevent distortion in graph embedding. *Advances in Neural Information Processing Systems*, 33:20779–20790, 2020.

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

[24] 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. 3

[25] Zhi Gao, Yuwei Wu, Yunde Jia, and Mehrtash Harandi. Curvature generation in curved spaces for few-shot learning. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 8691–8700, 2021. 3

[26] Alexandru Tifrea, Gary Bécigneul, and Octavian-Eugen Ganea. Poincar\'e glove: Hyperbolic word embeddings. *arXiv preprint arXiv:1810.06546*, 2018. 3

[27] 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. 3

[28] Ines Chami, Zhitao Ying, Christopher Ré, and Jure Leskovec. Hyperbolic graph convolutional neural networks. *Advances in neural information processing systems*, 32, 2019.

[29] Karan Desai, Maximilian Nickel, Tanmay Rajpurohit, Justin Johnson, and Shanmukha Ramakrishna Vedantam. Hyperbolic image-text representations. In *International Conference on Machine Learning*, pages 7694–7731. PMLR, 2023. 2, 3, 5

[30] Luke Vilnis and Andrew McCallum. Word representations via gaussian embedding. In *International Conference on Learning Representations*, 2015. 2

[31] Ben Athiwaratkun and Andrew Gordon Wilson. Multimodal word distributions. *arXiv preprint arXiv:1704.08424*, 2017. 3

[32] Ben Athiwaratkun and Andrew Gordon Wilson. Hierarchical density order embeddings. In *International Conference on Learning Representations*, 2018.

[33] Gengcong Yang, Jingyi Zhang, Yong Zhang, Baoyuan Wu, and Yujiu Yang. Probabilistic modeling of semantic ambiguity for scene graph generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12527–12536, 2021. 2

[34] 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. 2, 6

[35] Hugo Touvron, Matthieu Cord, Matthijs Douze, Francisco Massa, Alexandre Sablayrolles, and Hervé Jégou. Training data-efficient image transformers & distillation through attention. In *International conference on machine learning*, pages 10347–10357. PMLR, 2021. 2, 6, 7

[36] 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. 2, 6, 7, 8

[37] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V 13*, pages 740–755. Springer, 2014. 6, 7

[38] 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, 7

[39] Pedro F Felzenszwalb, Ross B Girshick, David McAllester, and Deva Ramanan. Object detection with discriminatively trained part-based models. *IEEE transactions on pattern analysis and machine intelligence*, 32(9):1627–1645, 2009. 2

[40] Feng Han and Song-Chun Zhu. Bottom-up/top-down image parsing with attribute grammar. *IEEE transactions on pattern analysis and machine intelligence*, 31(1):59–73, 2008.

[41] Erik B Sudderth, Antonio Torralba, William T Freeman, and Alan S Willsky. Learning hierarchical models of scenes, objects, and parts. In *Tenth IEEE International Conference on Computer Vision (ICCV'05) Volume 1*, volume 2, pages 1331–1338. IEEE, 2005.

[42] Zhuowen Tu, Xiangrong Chen, Alan L Yuille, and Song-Chun Zhu. Image parsing: Unifying segmentation, detection, and recognition. *International Journal of computer vision*, 63: 113–140, 2005. 2

[43] Tianfu Wu and Song-Chun Zhu. A numerical study of the bottom-up and top-down inference processes in and-or graphs. *International journal of computer vision*, 93:226–252, 2011. 2

[44] Wenguan Wang, Zhijie Zhang, Siyuan Qi, Jianbing Shen, Yanwei Pang, and Ling Shao. Learning compositional neural information fusion for human parsing. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 5703–5713, 2019. 2

[45] Wenguan Wang, Hailong Zhu, Jifeng Dai, Yanwei Pang, Jianbing Shen, and Ling Shao. Hierarchical human parsing with typed part-relation reasoning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 8929–8939, 2020. 2

[46] Sandro Braun, Patrick Esser, and Björn Ommer. Unsupervised part discovery by unsupervised disentanglement. In *Pattern Recognition: 42nd DAGM German Conference, DAGM GCPR 2020, Tübingen, Germany, September 28–October 1, 2020, Proceedings 42*, pages 345–359. Springer, 2021. 2

[47] Subhabrata Choudhury, Iro Laina, Christian Rupprecht, and Andrea Vedaldi. Unsupervised part discovery from contrastive reconstruction. *Advances in Neural Information Processing Systems*, 34:28104–28118, 2021.

[48] Wei-Chih Hung, Varun Jampani, Sifei Liu, Pavlo Molchanov, Ming-Hsuan Yang, and Jan Kautz. Scops: Self-supervised co-part segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 869–878, 2019. 2

[49] Tsung-Wei Ke, Sangwoo Mo, and Stella X. Yu. Learning hierarchical image segmentation for recognition and by recognition. In *The Twelfth International Conference on Learning Representations*, 2024. 2

[50] Sanghyuk Chun, Seong Joon Oh, Rafael Sampaio De Rezende, Yannis Kalantidis, and Diane Larlus. Probabilistic embeddings for cross-modal retrieval. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8415–8424, 2021. 3, 5

[51] Yichun Shi and Anil K Jain. Probabilistic face embeddings. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 6902–6911, 2019. 3

[52] Jungin Park, Jiyoung Lee, Ig-Jae Kim, and Kwanghoon Sohn. Probabilistic representations for video contrastive learning. In

Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pages 14711–14721, 2022. 3

[53] Maximillian Nickel and Douwe Kiela. Poincaré embeddings for learning hierarchical representations. *Advances in neural information processing systems*, 30, 2017. 3

[54] 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. 3

[55] Zhenzhen Weng, Mehmet Giray Ogut, Shai Limonchik, and Serena Yeung. Unsupervised discovery of the long-tail in instance segmentation using hierarchical self-supervision. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 2603–2612, 2021. 3

[56] Valentin Khrulkov, Leyla Mirvakhabova, Evgeniya Ustinova, Ivan Oseledets, and Victor Lempitsky. Hyperbolic image embeddings. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6418–6428, 2020. 3

[57] Durk P Kingma, Tim Salimans, and Max Welling. Variational dropout and the local reparameterization trick. *Advances in neural information processing systems*, 28, 2015. 4

[58] Aaron van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*, 2018. 5

[59] Mingxing Tan and Quoc Le. Efficientnet: Rethinking model scaling for convolutional neural networks. In *International conference on machine learning*, pages 6105–6114. PMLR, 2019. 6

[60] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 10012–10022, 2021. 6, 7

[61] Wenhai Wang, Enze Xie, Xiang Li, Deng-Ping Fan, Kaitao Song, Ding Liang, Tong Lu, Ping Luo, and Ling Shao. Pvt v2: Improved baselines with pyramid vision transformer. *Computational Visual Media*, 8(3):415–424, 2022. 6, 7

[62] Mingyu Ding, Bin Xiao, Noel Codella, Ping Luo, Jingdong Wang, and Lu Yuan. Davit: Dual attention vision transformers. In *European Conference on Computer Vision*, pages 74–92. Springer, 2022. 6

[63] Pengchuan Zhang, Xiyang Dai, Jianwei Yang, Bin Xiao, Lu Yuan, Lei Zhang, and Jianfeng Gao. Multi-scale vision longformer: A new vision transformer for high-resolution image encoding. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 2998–3008, 2021.

[64] Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. Focal loss for dense object detection. In *Proceedings of the IEEE international conference on computer vision*, pages 2980–2988, 2017. 6

[65] Elad Hoffer, Tal Ben-Nun, Itay Hubara, Niv Giladi, Torsten Hoefler, and Daniel Soudry. Augment your batch: Improving generalization through instance repetition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8129–8138, 2020. 6

[66] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017. 6

[67] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. Mask r-cnn. In *Proceedings of the IEEE international conference on computer vision*, pages 2961–2969, 2017. 7

[68] Yanghao Li, Hanzi Mao, Ross Girshick, and Kaiming He. Exploring plain vision transformer backbones for object detection. In *European Conference on Computer Vision*, pages 280–296. Springer, 2022. 7

[69] Alexander Kirillov, Ross Girshick, Kaiming He, and Piotr Dollár. Panoptic feature pyramid networks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 6399–6408, 2019. 7

[70] Tete Xiao, Yingcheng Liu, Bolei Zhou, Yuning Jiang, and Jian Sun. Unified perceptual parsing for scene understanding. In *Proceedings of the European conference on computer vision (ECCV)*, pages 418–434, 2018. 7