Hyperbolic Contrastive Learning for Visual Representations beyond Objects

Songwei Ge*[1], Shlok Mishra*[1], Simon Kornblith[2],
Chun-Liang Li[2], David Jacobs[1,3]
[1]University of Maryland, College Park, [2]Google Research, [3]Meta
{songweig,shlokm,dwj}@umd.edu, {chunliang,skornblith}@google.com

Abstract

Although self-/un-supervised methods have led to rapid progress in visual representation learning, these methods generally treat objects and scenes using the same lens. In this paper, we focus on learning representations for objects and scenes that preserve the structure among them. Motivated by the observation that visually similar objects are close in the representation space, we argue that the scenes and objects should instead follow a hierarchical structure based on their compositionality. To exploit such a structure, we propose a contrastive learning framework where a Euclidean loss is used to learn object representations and a hyperbolic loss is used to encourage representations of scenes to lie close to representations of their constituent objects in a hyperbolic space. This novel hyperbolic objective encourages the scene-object hypernymy among the representations by optimizing the magnitude of their norms. We show that when pretraining on the COCO and OpenImages datasets, the hyperbolic loss improves downstream performance of several baselines across multiple datasets and tasks, including image classification, object detection, and semantic segmentation. We also show that the properties of the learned representations allow us to solve various vision tasks that involve the interaction between scenes and objects in a zero-shot fashion.

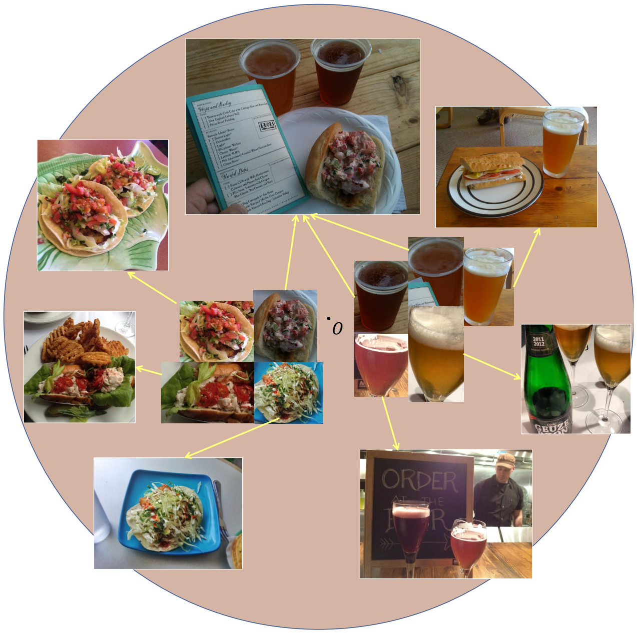

Figure 1. Illustration of the representation space learned by our models. Object images of the same class tend to gather near the center around similar directions, while the scene images are far away in these directions with larger norms.

1. Introduction

Our visual world is diverse and structured. Imagine taking a close-up of a box of cereal in the morning. If we zoom out slightly, we may see different nearby objects such as a pitcher of milk, a cup of hot coffee, today's newspaper, or reading glasses. Zooming out further, we will probably recognize that these items are placed on a dining table with the kitchen as background rather than inside a bathroom. Such scene-object structure is diverse, yet not completely random. In this paper, we aim at learning visual representations of both the cereal box (objects) and the entire dining table (scenes) in the same space while preserving such hierarchical structures.

Un-/self-supervised learning has become a standard method to learn visual representations [7, 12, 24, 26, 27, 51]. Although these methods attain superior performance over supervised pretraining on object-centric datasets such as ImageNet [6], inferior results are observed on images depicting multiple objects such as OpenImages or COCO [68]. Several methods have been proposed to mitigate this issue, but all focus either on learning improved object representations [1, 68] or dense pixel representations [39, 64, 69], instead of explicitly modeling representations for scene images. The object representations learned by these methods present a natural topology [67]. That is, the objects from visually similar

*Equal Contribution. The order is decided randomly.

classes lie close to each other in the representation space. However, it is not clear how the representations of scene images should fit into that topology. Directly applying existing contrastive learning results in a sub-optimal topology of scenes and objects as well as unsatisfactory performance, as we will show in the experiments. To this end, we argue that a hierarchical structure can be naturally adopted. Considering that the same class of objects can be placed in different scenes, we construct a hierarchical structure to describe such relationships, where the root nodes are the visually similar objects, and the scene images consisting of them are placed as the descendants. We call this structure the object-centric scene hierarchy.

The intermediate modeling difficulty induced by this structure is the combinatorial explosion. A finite number of objects leads to exponentially many different possible scenes. Consequently, Euclidean space may require an arbitrarily large number of dimensions to faithfully embed these scenes, whereas it is known that any infinite trees can be embedded without distortion in a 2D hyperbolic space [25]. Therefore, we propose to employ a hyperbolic objective to regularize the scene representations. To learn representations of scenes, in the general setting of contrastive learning, we sample co-occurring scene-object pairs as positive pairs, and objects that are not part of that scene as negative samples, and use these pairs to compute an auxiliary hyperbolic contrastive objective. Our model is trained to reduce the distance between positive pairs and push away the negative pairs in a hyperbolic space.

Contrastive learning usually has objectives defined on a hypersphere [12, 27]. By discarding the norm information, these models circumvent the shortcut of minimizing losses through tuning the norms and obtain better downstream performance. However, the norm of the representation can also be used to encode useful representational structure. In hyperbolic space, the magnitude of a vector often plays the role of modeling the hypernymy of the hierarchical structure [45, 53, 59]. When projecting the representations to the hyperbolic space, the norm information is preserved and used to determine the Riemannian distance, which eventually affects the loss. Since hyperbolic space is diffeomorphic and conformal to Euclidean space, our hyperbolic contrastive loss is differentiable and complementary to the original contrastive objective.

When training simultaneously with the original contrastive objective for objects and our proposed hyperbolic contrastive objective for scenes, the resulting representation space exhibits a desired hierarchical structure while leaving the object clustering topology intact as shown in Figure 1. We demonstrate the effectiveness of the hyperbolic objective under several frameworks on multiple downstream tasks. We also show that the properties of the representations allow us to perform various vision tasks in a zero-shot way, from label uncertainty quantification to out-of-context object detection. Our contributions are summarized below:

1. We propose a hyperbolic contrastive loss that regularizes scene representations so that they follow an object-centric hierarchy, with positive and negative pairs sampled from the hierarchy.

2. We demonstrate that our learned representations transfer better than representations learned using vanilla contrastive loss on a variety of downstream tasks, including object detection, semantic segmentation, and linear classification.

3. We show that the magnitude of representation norms effectively reflect the scene-objective hypernymy.

2. Method

In this section, we elaborate upon our approach to learning visual representations of object and scene images. We start by describing the hierarchical structure between objects and scenes that we wish to enforce in the learned representation space.

2.1. Object-Centric Scene Hierarchy

From simple object co-occurrence statistics [19, 41] to finer object relationships [30, 32], using hierarchical relationships between objects and scenes to understand images is not new. Previous studies primarily work on an image-level hierarchy by dividing an image into its lower-level elements recursively: a scene contains multiple objects, an object has different parts, and each part may consist of even lower-level features [14, 29, 48]. While this is intuitive, it describes a hierarchical structure contained in the individual images. Instead, we study the structure presented among different images. Our goal is to learn a representation space for images of both objects and scenes across the entire dataset. To this end, we argue that it is more natural to consider an *object-centric hierarchy*.

It is known that when training an image classifier, the objects from visually similar classes often lie close to each other in the representation space [67], which has become the cornerstone of contrastive learning. Motivated by this observation, we believe that the representation of each scene image should also be close to the object clusters it consists of. However, modeling scenes requires a much larger volume due to the exponential number of possible compositions of objects. Another way to think about the object-centric hierarchy is through the generality and specificity as often discussed in the language literature [42, 45]. An object concept is general when standing alone in the visual world, and it will become specific when a certain context is given. For example, “a desk” is thought to be a more general concept than “a desk in a classroom with a boy sitting on it”.

Therefore, we propose to study an object-centric hierarchy across the entire dataset. Formally, given a set of images $\mathcal{S} = \{s_1, s_2, \cdots, s_n\}$, $\mathcal{O}_i = \{o_i^1, o_i^2, \cdots, o_i^{n_i}\}$ are the object bounding boxes contained in the image s_i. We define the regions of scene $\mathcal{R}_i = \{r_i^1, r_i^2, \cdots, r_i^{m_i}\}$ to be partial areas of the image s_i that contain multiple objects such that $r_i^j = \cup_k o_i^k$, where $o_i^k \in \mathcal{O}_i$ and object k is in the region j. We define the object-centric hierarchy $T = (V, E)$ to be that $V = \mathcal{S} \cup \mathcal{O} \cup \mathcal{R}$, where $\mathcal{R} = \mathcal{R}_1 \cup \cdots \cup \mathcal{R}_n$ and $\mathcal{O} = \mathcal{O}_1 \cup \cdots \cup \mathcal{O}_n$. For $u, v \in V$, $e = (u, v)$ is an edge of T if $u \subseteq v$ or $v \subseteq u$. Note that the natural scene images $\mathcal{S}$ are always put as the leaf nodes.

2.2. Representation Learning beyond Objects

To describe our proposed model based on this hierarchy, we begin with a brief review of hyperbolic space and its properties used in our model. For comprehensive introductions to Riemannian geometry and hyperbolic space, we refer the readers to [16, 34].

2.2.1 Hyperbolic Space

A hyperbolic space $(\mathbb{H}^m, g)$ is a complete, connected Riemannian manifold with constant negative sectional curvature. These special manifolds are all isometric to each other with the isometries defined as $O^+(m, 1)$. Among these isometries, there are five common models that previous studies often work on [5]. In this paper, we choose the Poincaré ball $\mathbb{D}^n := \left\{p \in \mathbb{R}^n \mid \|p\|^2 < r^2\right\}$ as our basic model [21, 45, 59], where $r > 0$ is the radius of the ball. The Poincaré ball is coupled with a Riemannian metric $g_{\mathbb{D}}(p) = \frac{4}{(1-\|p\|^2/r^2)^2} g_{\mathbb{E}}$, where $p \in \mathbb{D}^n$ and $g_{\mathbb{E}}$ is the canonical metric of the Euclidean space. For $p, q \in \mathbb{D}$, the Riemannian distance on the Poincaré ball induced by its metric $g_{\mathbb{D}}$ is defined as follows:

$$d_{\mathbb{D}}(p, q) = 2r \tanh^{-1}\left(\frac{\|-p \oplus q\|}{r}\right), \tag{1}$$

where $\oplus$ is the Möbius addition and it is clearly differentiable. In addition, the Poincaré ball can be viewed as a natural counterpart of the hypersphere as it allows all directions, unlike the other models such as the halfspace or hemisphere models that have constraints on the directions. The hyperbolic space is globally differomorphic to the Euclidean space, which is stated in the theorem below:

Theorem 1. ***(Cartan–Hadamard).*** *For every point $p \in \mathbb{H}^n$ the exponential map $\exp_p : T_p\mathbb{H}^n \approx \mathbb{R}^n \to \mathbb{H}^n$ is a smooth covering map. Since $\mathbb{H}^n$ is simply connected, it is diffeomorphic to $\mathbb{R}^n$.*

Specifically, for $p \in \mathbb{D}^n$ and $v \in T_p\mathbb{D}^n \approx \mathbb{R}^n$, the exponential map of the Poincaré ball $\exp_p : T_p\mathbb{D}^n \to \mathbb{D}^n$ is defined as

$$\exp_p(v) := p \oplus \left(\tanh\left(\frac{r\|v\|}{r^2 - \|p\|^2}\right)\frac{rv}{\|v\|}\right), \tag{2}$$

The exponential map gives us a way to map the output of a network, which is in the Euclidean space, to the Poincaré ball. In practice, to avoid numerical issues, we clip the maximal norm of v with $r - \varepsilon$ before the projection, where $\varepsilon > 0$. During the backpropagation, we perform RSGD [4] by scaling the gradients by $g_{\mathbb{D}}(p)^{-1}$. Intuitively, this forces the optimizer to take a smaller step when p is closer to the boundary. The scaling factor is lower bounded by $\mathcal{O}(\varepsilon^2)$.

The immediate consequence of the negative curvature is that for any point $\boldsymbol{p} \in \mathbb{H}^m$, there are no conjugate points along any geodesic starting from $\boldsymbol{p}$. Therefore, the volume grows exponentially faster in hyperbolic space than in Euclidean space. Such a property makes it suitable to embed the hierarchical structure that has constant branching factors and exponential number of nodes. This is formally stated in the theorem below:

Theorem 2. *[25] Given a Poincaré ball $\mathbb{D}^n$ with an arbitrary dimension $n \geq 2$ and any set of points $p_1, \cdots, p_m \in \mathbb{D}^n$, there exists a finite weighted tree (T, d_T) and an embedding $f : T \to \mathbb{D}^n$ such that for all i, j,*

$$\left|d_T\left(f^{-1}(x_i), f^{-1}(x_j)\right) - d_{\mathbb{D}}(x_i, x_j)\right| = \mathcal{O}(\log(1+\sqrt{2})\log(m))$$

Intuitively, the theorem states that any tree can be embedded into a Poincaré disk ($n = 2$) with low distortion. On the contrary, it is known that the Euclidean space with unbounded number of dimensions is not able to achieve such a low distortion [36]. One useful intuition [53] to help understand the advantage of the hyperbolic space is given two points $p, q \in \mathbb{D}^n$ s.t. $\|p\| = \|q\|$,

$$d_{\mathbb{D}}(p, q) \to d_{\mathbb{D}}(p, 0) + d_{\mathbb{D}}(0, q), \text{ as } \|p\| = \|q\| \to r \tag{3}$$

This property basically reflects the fact that the shortest path in a tree is the path through the earliest common ancestor, and it is reproduced in the Poincaré when points are both close to the boundary.

2.2.2 Hyperbolic Contrastive Learning

Given the theoretical benefits of the hyperbolic space stated above, we propose a contrastive learning framework as shown in Figure 2. We adopt two losses to learn the object and scene representations. First, to learn object representations, we use the standard normalized temperature-scaled cross-entropy loss, which operates on the hypersphere in Euclidean space. As shown in the top branch of Figure 2, we crop two views of a jittered and slightly expanded object region as the positive pairs and feed into the base and momentum encoders to calculate the object representations. We

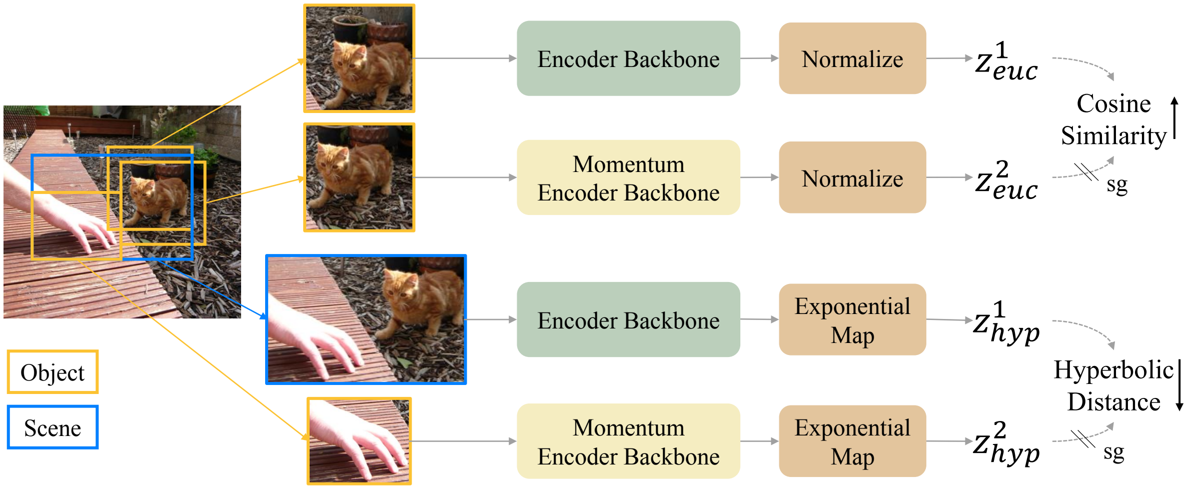

Figure 2. Our **H**yperbolic **C**ontrastive **L**earning (HCL) framework has two branches: given a scene image, two object regions are cropped to learn the object representations with a loss defined in the Euclidean space focusing on the representation directions. A scene region as well as a contained object region are used to learn the scene representations with a loss defined in the hyperbolic space that affects the representation norms.

denote the output after the normalization to be $\mathbf{z}^1_{\text{euc}}$ and $\mathbf{z}^2_{\text{euc}}$. We follow MoCo [27] and leverage a memory bank to store the negative representations z^n_{euc}, which are the features $\mathbf{z}^2_{\text{euc}}$ from the previous batches. Note that our framework can be readily extended to other contrastive learning models. The Euclidean loss for each image is then calculated as:

$$\mathcal{L}_{\text{euc}} = -\log \frac{\exp\left(\mathbf{z}^1_{\text{euc}} \cdot \mathbf{z}^2_{\text{euc}}/\tau\right)}{\exp\left(\mathbf{z}^1_{\text{euc}} \cdot \mathbf{z}^2_{\text{euc}}/\tau\right) + \sum_n \exp\left(\mathbf{z}^1_{\text{euc}} \cdot \mathbf{z}^n_{\text{euc}}/\tau\right)},$$

where τ is a temperature parameter.

While the loss above aims to learn object representations, we propose a hyperbolic contrastive objective to learn the representations for scene images. We sample positive region pairs u and v from object-centric scene hierarchy T such that $(u, v) \in E$. In other words, as shown in the bottom branch of Figure 2, the objects contained in one region are required to be a subset of the objects in the other. We sample the negative samples of u to be $\mathcal{N}_u = \{v|(u, v) \notin E\}$. However, building and sampling exhaustively from the entire hierarchy explicitly is tricky. In practice, given an image s, we always sample $u \in \mathcal{R} \cup \{s\}$ to be a scene region, $v \in \mathcal{O}$ to be an object that occurs in u, and $\mathcal{N}_u$ to be the other objects that are not in u.

The pair of scene and object images are fed into the base and momentum encoders that share the weights with the Euclidean branch. However, instead of normalizing the output of the encoders, we use the exponential map defined in the equation 2 to project these features in the Euclidean space to the Poincaré ball, which are denoted as $\mathbf{z}^1_{\text{hyp}}$ and $\mathbf{z}^2_{\text{hyp}}$. Further, we replace the inner product in the cross-entropy loss with the negative hyperbolic distance as defined in equation 1. We calculate the hyperbolic contrastive loss as follows:

$$\mathcal{L}_{\text{hyp}} = -\log \frac{\exp\left(-\frac{d_{\mathbb{D}}(\mathbf{z}^1_{\text{hyp}}, \mathbf{z}^2_{\text{hyp}})}{\tau}\right)}{\exp\left(-\frac{d_{\mathbb{D}}(\mathbf{z}^1_{\text{hyp}}, \mathbf{z}^2_{\text{hyp}})}{\tau}\right) + \sum_n \exp\left(-\frac{d_{\mathbb{D}}(\mathbf{z}^1_{\text{hyp}}, \mathbf{z}^n_{\text{hyp}})}{\tau}\right)},$$

When minimizing the distances of all the positive pairs, with the intuition from equation 3, it would be beneficial to put the nodes near the root, i.e. objects, close to the center to achieve an overall lower loss. The overall loss function of our model is as follows:

$$\mathcal{L} = \mathcal{L}_{\text{euc}} + \lambda \mathcal{L}_{\text{hyp}},$$

where λ is a scaling parameter to control the trade-off between hyperbolic and Euclidean losses.

3. Experiments

3.1. Implementation Details

Pre-training phase. We pre-train on three datasets: COCO [35], the full OpenImages labelled dataset [33]($\sim$ 1.7 million samples) and a subset of OpenImages ($\sim 212k$) [44]. All these datasets are multi-object datasets; OpenImages contains 12 objects on average per image and COCO contains 6 objects on average. We experiment with both the ground truth bounding box (GT) and using selective search (SS) [61] to produce object bounding boxes in an unsupervised fashion, following previous work [68]. As the goal of this paper is not to present another state-of-the-art self-supervised learning method, we implement our sampling

	AP^{b}	$\mathrm{AP}^{\mathrm{b}}_{50}$	$\mathrm{AP}^{\mathrm{b}}_{75}$	AP^{m}	$\mathrm{AP}^{\mathrm{m}}_{50}$	$\mathrm{AP}^{\mathrm{m}}_{75}$
MoCo-v2 pre-trained on COCO:						
Baseline	38.5	58.1	42.1	34.8	55.3	37.3
HCL w/o $\mathcal{L}_{\mathrm{hyp}}$	39.7	60.1	43.4	36.0	57.3	38.8
HCL CC	**40.6**	**61.1**	**44.5**	**37.0**	**58.3**	**39.7**
Dense-CL pre-trained on COCO:						
Baseline	39.6	59.3	43.3	35.7	56.5	38.4
HCL w/o $\mathcal{L}_{\mathrm{hyp}}$	41.3	61.5	44.7	37.5	59.5	40.4
HCL	**42.5**	**62.5**	**45.8**	**38.5**	**60.6**	**41.4**
ORL pre-trained on COCO:						
Baseline	40.3	60.2	44.4	36.3	57.3	38.9
HCL	**41.4**	**61.4**	**45.5**	**37.3**	**58.5**	**40.0**
Dense-CL pre-trained on OpenImages:						
Baseline	38.2	58.9	42.6	34.8	55.3	37.8
HCL w/o $\mathcal{L}_{\mathrm{hyp}}$	41.1	61.5	44.4	37.2	58.3	39.7
HCL	**42.1**	**62.6**	**45.5**	**38.3**	**59.4**	**40.6**

Table 1. **Comparison with state-of-the-art methods.** This table shows object detection (columns 1-3) and semantic segmentation (columns 4-6) results on COCO using MoCo-v2, Dense-CL and ORL by pre-training on COCO and OpenImages using unsupervised object bounding boxes generated by the selective search. The first row in each sub-table shows the results using random crops on pre-training datasets. The second and third rows set HCL/$\mathcal{L}_{\mathrm{hyp}}$ to 0, which means we are pre-training baseline methods on just proposal boxes. Our model consistently improves both object detection and semantic segmentation tasks across multiple contrastive learning baselines by pre-training on both COCO (800 epochs) and the full OpenImages dataset (75 epochs, last 3 rows).

procedure and hyperbolic loss on top of three popular contrastive learning methods: MoCo-v2 [13], Dense-CL [64], and ORL [68]. Dense-CL is a contrastive learning framework which extracts dense features from scene images and generally achieves better object detection results than MoCo-v2. ORL is a pipeline that learns improved object representations from scene images. We also consider HCL without the hyperbolic loss $\mathcal{L}_{\mathrm{hyp}}$. This approach, which we denote as "HCL w/o $\mathcal{L}_{\mathrm{hyp}}$", adopts the same cropping strategy as HCL but applies only a standard contrastive loss. We show that adding the hyperbolic loss improves results under various settings. More details on the datasets as well as training setups can be found in Appendix A.

Downstream tasks. We evaluate our pre-trained models on image classification, object-detection and semantic segmentation. For classification, we show linear evaluation (lineval) accuracy with MoCo-v2, i.e. we freeze the backbone and only train the final linear layer. We test on VOC [18], ImageNet-100 [58] and ImageNet-1k [15] datasets. For object detection and semantic segmentation, we show results with all 3 baselines on the COCO datasets using Mask R-CNN, following [13]. We closely follow the common protocols listed in Detectron2 [66].

	Pre-train	Bbox	VOC	IN-100	IN-1k
MoCo-v2	COCO	-	64.79	64.84	51.17
HCL w/o $\mathcal{L}_{\mathrm{hyp}}$	COCO	SS	73.13	73.84	54.21
HCL w/o $\mathcal{L}_{\mathrm{hyp}}$	COCO	GT	75.55	76.22	54.52
HCL	COCO	SS	74.19	75.16	55.03
HCL	COCO	GT	**76.51**	**76.74**	**55.63**
MoCo-v2	OpenImages	-	69.95	72.80	54.12
HCL w/o $\mathcal{L}_{\mathrm{hyp}}$	OpenImages	SS	71.82	75.33	56.58
HCL w/o $\mathcal{L}_{\mathrm{hyp}}$	OpenImages	GT	73.79	77.36	57.57
HCL	OpenImages	SS	74.31	78.14	58.12
HCL	OpenImages	GT	**75.40**	**79.08**	**58.51**

Table 2. **Classification results with linear evaluation.** The first row shows the results using random crops on pre-training datasets. In the last two rows we use our hyperbolic loss and we see improved performance by using both Ground Truth (GT) boxes and Selective Search (SS) boxes. HCL improves scene-level classification on the VOC dataset, and object-level classification on ImageNet-100 and ImageNet-1k datasets.

3.2. Main Results

Object detection and semantic segmentation. Table 1 reports the object detection and semantic segmentation results by pre-training on COCO and full OpenImages dataset (last 3 rows) by using selective search boxes. HCL shows consistent improvements over the baselines on COCO object detection and COCO semantic segmentation. Although Dense-CL and ORL improve the object-level downstream performance over MoCo-v2 through improved object representations or dense pixel representations, they still lack the direct modeling of scene images. We show that learning representations for scene images in hyperbolic space is beneficial to object-level downstream performance. Note that pre-training Dense-CL on ImageNet for 200 epochs gives 40.3 mAP [64], while pre-trainng on OpenImages for only 75 epochs with our method gives 42.1 mAP. This shows the importance of efficient pre-training on datasets like OpenImages.

Image classification. As shown in Table 2, HCL improves image classification on both scene-level (VOC) and object-level (ImageNet) datasets. When pretraining on OpenImages, HCL improves ImageNet lineval accuracy by 0.94% points and VOC lineval classification accuracy by 1.61 mAP. We observe similar improvements when pretraining on COCO. HCL improves accuracy whether we use ground truth object bounding boxes or boxes generated by selective search. In general, we observe a larger improvement of using HCL on OpenImages than COCO, which supports our hypothesis that HCL provides larger improvements on datasets with

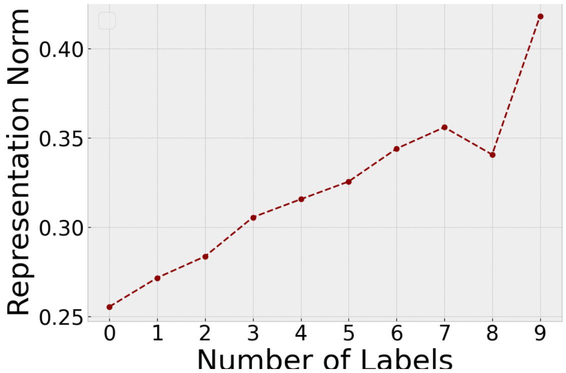

Figure 3. Average representation norms of images with different number of labels in ImageNet-ReaL.

Method	Indicator	Datasets	
		IN-Real	COCO
MoCo	Entropy	0.633	0.791
Supervised	Entropy	0.671	0.793
HCL	Norm	0.655	**0.839**
Ensemble	Entropy+Norm	**0.717**	0.823

Table 3. NDCG scores of the image rankings based on the different indicators and models, and evaluated by the number of labels per image.

more objects per image.

3.3. Properties of Models Trained with HCL

The visual representations learned by HCL have several useful properties. In this section, we evaluate the representation norm as an measure of the label uncertainty for image classification datasets, and evaluate the object-scene similarity in terms of out-of-context detection.

3.3.1 Label Uncertainty Quantification

ImageNet [15] is an image classification dataset consisting of object-centered images, each of which has a single label. As the performance on this dataset has gradually saturated, the original labels have been scrutinized more carefully [3, 52, 55, 60, 62]. Prevailing labeling issues in the validation set have been recently identified, including labeling errors, multi-label images with only a single label provided, and so on. Although [3] provides reassessed labels for the entire validation set, relabeling the entire training set may be infeasible.

Our learned representations provide a potential automatic way to identify images with multiple labels from datasets like ImageNet. Specifically, we first show in Figure 3 that there is a strong correlation between the representation norms and the number of labels per image according to the reassessed labels. For each class of the ImageNet training set, we use a pre-trained OpenImages model and rank the images according to their norms. The extreme images of some classes are shown in Figure 4 and also in the Appendix. Images with smaller norms tend to capture a single object, while those with larger norms are likely to depict a scene.

To quantitatively evaluate this property, we report the NDCG metric on the ranked images as shown in Table 3. NDCG assesses how often the scene images are ranked at the top. As a baseline, we rank the images based on the entropy of the class probability predicted by a classifier, which is a widely adopted indicator of label uncertainty [11, 47]. We use both MoCo-v2 and supervised ResNet-50 as the classifier. As shown in Table 3, using norms with HCL achieves similar rank quality as using entropy with the supervised ResNet-50 on the ImageNet-ReaL dataset. In addition, when combining two ranks using simple ensemble methods such as Borda count, the score is further improved to 0.717. This shows that the entropy and the norm provide complimentary signals regarding the existence of multiple labels. For example, the entropy indicator can be affected by the bias of the model and the norm indicator can be wrong on the images with multiple objects from the same class.

Compared to supervised indicators of label uncertainty, HCL has the additional advantage that it is dataset-agnostic and can be applied to new data without further training. To demonstrate this benefit, we report the same metric on the COCO validation, where we also have the number of labels for each image. Our method achieves much better NDCG scores than the supervised ResNet-50 as shown in Table 3. This finding can be potentially useful to guide label reassessment, or provide an extra signal for model training.

3.3.2 Out-of-Context Detection

Our hyperbolic loss $\mathcal{L}_{\text{hyp}}$ encourages the model to capture the similarity between the object and scene. We apply the resulting representations to detect out-of-context objects, which can be useful in designing data augmentation for object detection [17]. We are especially interested in out-of-context images with conflicting backgrounds. To this end, we use the out-of-context images proposed in the SUN09 dataset [14]. We first compute the representations of each object and entire scene image with that object masked out. We then calculate the hyperbolic distance between the representations mapped to the Poincaré ball. Some example images from this dataset as well as the distance of each contained object are shown in Figure 5. We find that the out-of-context objects generally have a large distance, i.e. smaller similarity, to the overall scene image. To quantify this finding, we compute the mAP of the object ranking on each image and obtain 0.61 for HCL. As a comparison, the MoCo similarity gives mAP $= 0.52$ and the random ranking gives mAP $= 0.44$.

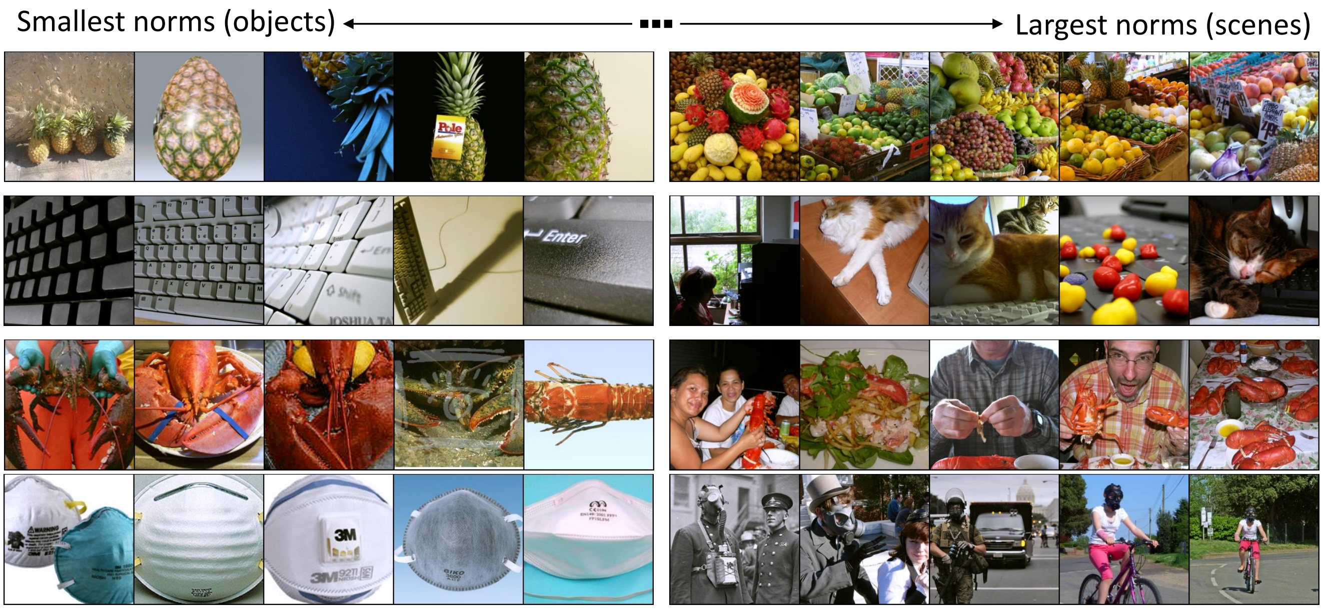

Figure 4. Images from ImageNet training set. The 5 images on the left have the smallest representation norms among all the images from the same class, and the 5 on the right have the largest norms.

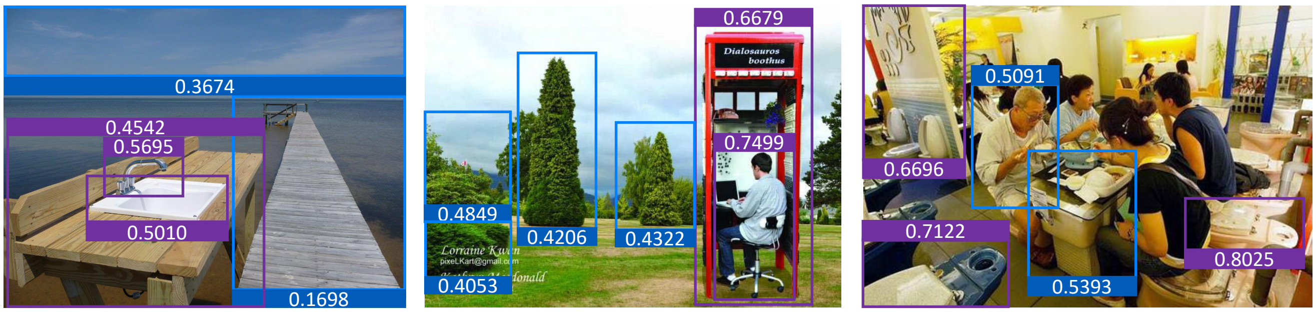

Figure 5. Out-of-context images from the SUN09 dataset. The bounding box of each object and its hyperbolic distance to the scene are shown. Regular objects are in blue and out-of-context objects are in purple. Note that the out-of-context objects tend to have large distances.

4. Main Ablation Studies

In this section, we report the results of several important ablation studies with respect to HCL. All the models are trained on the subset of the OpenImages dataset and linearly evaluated on the ImageNet-100 dataset. The top-1 accuracy is reported.

Similarity measure and the center of the scene-object hierarchy. We propose to use the negative hyperbolic distance as the similarity measure of the scene-object pairs. As an alternative, one can use cosine similarity on the hypersphere as the measure as in the original contrastive objective. However, this would attempt to maximize the similarity between a single object and multiple objects. It is likely that these objects belong to different classes, and hence this strategy impairs the quality of the representation. As shown in Table 4, replacing the negative hyperbolic distance with the Euclidean similarity impairs downstream performance. The resulting model performs even worse than the baseline without loss function on the scene-object pairs, demonstrating the necessity of using hyperbolic distance. We also validate our choice of an object-centric hierarchy by comparing its performance with that of a scene-centric hierarchy [48, 49] generated by sampling the negative pairs as objects and unpaired scenes. This scene-centric hierarchy leads to substantially lower accuracy (Table 4).

Trade-off between the Euclidean and hyperbolic losses. We adopt the Euclidean loss to learn object-object similarity and the hyperbolic loss to learn object-scene similarity. A hyperparameter λ controls the trade-off between them. As shown in Table 5, we find that a smaller $\lambda = 0.01$ leads to marginal improvement. However, we also observe that larger λs can lead to unstable and even stalled training. With careful inspection, we find that in the early stage of the training, the gradient provided by the hyperbolic loss can be inaccurate but strong, which pushes the representations to be

Distance	Center	IN-100 Accuracy
-	-	77.36
Hyperbolic	Scene	79.08
Hyperbolic	Object	76.96
Euclidean	Scene	76.68

Table 4. Similarity measure and hierarchy center.

λ	IN-100 Accuracy
0.01	77.70
0.1	79.08
0.2	78.64
0.5	0

Table 5. Losses trade-off.

Optimizer	λ	IN-100 Accuracy
RSGD	0.1	79.08
RSGD	0.5	0
SGD	0.1	70.16
SGD	0.5	74.18

Table 6. RSGD versus SGD optimizers.

close to the boundary. As a result, since Riemannian SGD divides gradients by the distance to the boundary, updates become small and training ceases to make progress.

Optimizer. Given the observation above, we ask whether RSGD is necessary for practical usage. We replace the RSGD optimizer with SGD. To avoid numerical issues when the representations are too close to the boundary, we increase ε from $1e^{-5}$ to $1e^{-1}$. This allows a larger λ to be used as opposed to the RSGD. However, SGD always yields inferior performance compared to RSGD.

5. Related Work

Representation Learning with Hyperbolic Space. Representations are typically learned in Euclidean space. Hyperbolic space has been adopted for its expressiveness in modeling tree-like structures existing in various domains such as language [45, 46, 53], graphs [2, 8, 50], and vision [10, 57]. The corresponding neural network modules have been designed to boost the progress of such applications [9, 21, 37, 56]. The hierarchical structure presented in the datasets can arise from three factors that motivate the use of hyperbolic space. The first factor is generality: the hypernym-hyponym property is a natural feature of words (e.g. WordNet [42]) and the hyperbolic space is extensively exploited to learn word and image embeddings that preserve that property [20, 38, 40, 53, 59, 70]. The second factor is uncertainty: Several studies have found that applying hyperbolic neural network modules to different tasks leads to a natural modeling of the uncertainty [23,31,57]. The third factor is compositionality of different basic elements to form a natural hierarchy. Motivated by these factors, previous work in computer vision has applied hierarchical representations learned in the hyperbolic space to various tasks such as image classification [31] or segmentation [65], zero-/few-shot learning [38], action recognition [40], and video prediction [57]. In this paper, we focus on learning the representations that capture the hierarchy between the objects and scenes with the goal of learning general-purpose image representations that can transfer to various downstream tasks.

Self-Supervised Learning on Scenes. Self-Supervised Learning (SSL) has made great strides in closing the performance with supervised methods [12, 13, 22] when pretrained on the object-centric datasets like ImageNet. However, recent work has shown that SSL is limited on multi-object datasets like COCO [43, 54, 64] and OpenImages [33]. Several papers mitigate this issue by proposing different techniques. Dense-CL [64] operates on pre-average pool features and uses dense features on pixel level to show improved performance on dense tasks such as semantic segmentation. Det-Con [28] uses unsupervised semantic segmentation masks to generate features for the corresponding objects in the two views. PixContrast [69] uses pixel-to-propagation consistency pretext task to build features for both dense downstream tasks and discriminative downstream tasks. Pixel-to-Pixel Contrast [63] uses pixel-level contrastive learning to learn better features for semantic segmentation. Self-EMD [39] uses earth mover distance with BYOL [24] for pretraining on the COCO dataset. ORL [68] uses selective search to generate object proposals, then applies object-level contrastive loss to enforce object-level consistency. Below-par performance of SSL methods can be attributed to treating scenes and objects using similar techniques, which often results in similar representations. In our work, instead of treating scenes and objects similarly, we use a hyperbolic loss, which builds representation that disambiguates scenes and objects based on the norm of the embeddings. Our method not only separates scenes and objects, but also improves downstream tasks such as image classification.

6. Conclusion

We present HCL, a contrastive learning framework that learns visual representation for both objects and scenes in the same representation space. The major novelty of our method is a hyperbolic contrastive objective built on an object-centric scene hierarchy. We show the effectiveness of HCL on several benchmarks including image classification, object detection, and semantic segmentation. We also demonstrate useful properties of the representations under several zero-shot settings, from detecting out-of-context objects to quantifying the label uncertainty in the datasets like ImageNet. More generally, we hope this paper will encourage future work towards building a more holistic visual representation space, and draw attention to the power of non-Euclidean representation learning.

7. Acknowledgements

Songwei Ge, Shlok Mishra, and David Jacobs were supported in part by the National Science Foundation under grant no. IIS-1910132 and IIS-2213335.

References

[1] Yutong Bai, Xinlei Chen, Alexander Kirillov, Alan Yuille, and Alexander C Berg. Point-level region contrast for object detection pre-training. *CVPR*, 2022. 1

[2] Ivana Balazevic, Carl Allen, and Timothy Hospedales. Multi-relational poincaré graph embeddings. *NeurIPS*, 2019. 8

[3] Lucas Beyer, Olivier J. Hénaff, Alexander Kolesnikov, Xiaohua Zhai, and Aäron van den Oord. Are we done with imagenet?, 2020. 6

[4] Silvere Bonnabel. Stochastic gradient descent on riemannian manifolds. *IEEE Transactions on Automatic Control*, 58(9):2217–2229, 2013. 3

[5] James W Cannon, William J Floyd, Richard Kenyon, Walter R Parry, et al. Hyperbolic geometry. *Flavors of geometry*, 31(59-115):2, 1997. 3

[6] Mathilde Caron, Ishan Misra, Julien Mairal, Priya Goyal, Piotr Bojanowski, and Armand Joulin. Unsupervised learning of visual features by contrasting cluster assignments. *NeurIPS*, 2020. 1

[7] Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers. In *ICCV*, 2021. 1

[8] Ines Chami, Adva Wolf, Da-Cheng Juan, Frederic Sala, Sujith Ravi, and Christopher Ré. Low-dimensional hyperbolic knowledge graph embeddings. In *ACL*, 2020. 8

[9] Ines Chami, Zhitao Ying, Christopher Ré, and Jure Leskovec. Hyperbolic graph convolutional neural networks. *NeurIPS*, 2019. 8

[10] Jiaxin Chen, Jie Qin, Yuming Shen, Li Liu, Fan Zhu, and Ling Shao. Learning attentive and hierarchical representations for 3d shape recognition. In *ECCV*, 2020. 8

[11] Pengfei Chen, Ben Ben Liao, Guangyong Chen, and Shengyu Zhang. Understanding and utilizing deep neural networks trained with noisy labels. In *ICML*, 2019. 6

[12] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *ICML*, 2020. 1, 2, 8

[13] Xinlei Chen, Haoqi Fan, Ross Girshick, and Kaiming He. Improved baselines with momentum contrastive learning. *arXiv preprint arXiv:2003.04297*, 2020. 5, 8

[14] Myung Jin Choi, Joseph J Lim, Antonio Torralba, and Alan S Willsky. Exploiting hierarchical context on a large database of object categories. In *CVPR*, 2010. 2, 6

[15] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *CVPR*, 2009. 5, 6

[16] Manfredo Perdigao Do Carmo and J Flaherty Francis. *Riemannian geometry*, volume 6. Springer, 1992. 3

[17] Nikita Dvornik, Julien Mairal, and Cordelia Schmid. On the importance of visual context for data augmentation in scene understanding. *PAMI*, 43(6):2014–2028, 2019. 6

[18] Mark Everingham, Luc Van Gool, Christopher KI Williams, John Winn, and Andrew Zisserman. The pascal visual object classes (voc) challenge. *IJCV*, 88(2):303–338, 2010. 5

[19] Carolina Galleguillos, Andrew Rabinovich, and Serge Belongie. Object categorization using co-occurrence, location and appearance. In *CVPR*, 2008. 2

[20] Octavian Ganea, Gary Bécigneul, and Thomas Hofmann. Hyperbolic entailment cones for learning hierarchical embeddings. In *ICML*, 2018. 8

[21] Octavian Ganea, Gary Bécigneul, and Thomas Hofmann. Hyperbolic neural networks. *NeurIPS*, 2018. 3, 8

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

[23] Mina GhadimiAtigh, Julian Schoep, Erman Acar, Nanne van Noord, and Pascal Mettes. Hyperbolic image segmentation. *arXiv preprint arXiv:2203.05898*, 2022. 8

[24] Jean-Bastien Grill, Florian Strub, Florent Altché, Corentin Tallec, Pierre Richemond, Elena Buchatskaya, Carl Doersch, Bernardo Avila Pires, Zhaohan Guo, Mohammad Gheshlaghi Azar, Bilal Piot, koray kavukcuoglu, Remi Munos, and Michal Valko. Bootstrap your own latent - a new approach to self-supervised learning. In *NeurIPS*, 2020. 1, 8

[25] Mikhael Gromov. Hyperbolic groups. In *Essays in group theory*, pages 75–263. Springer, 1987. 2, 3

[26] Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked autoencoders are scalable vision learners. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 16000–16009, June 2022. 1

[27] Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. Momentum contrast for unsupervised visual representation learning. In *CVPR*, 2020. 1, 2, 4

[28] Olivier J Hénaff, Skanda Koppula, Jean-Baptiste Alayrac, Aaron van den Oord, Oriol Vinyals, and João Carreira. Efficient visual pretraining with contrastive detection. In *ICCV*, pages 10086–10096, 2021. 8

[29] Geoffrey Hinton. How to represent part-whole hierarchies in a neural network. *arXiv preprint arXiv:2102.12627*, 2021. 2

[30] Justin Johnson, Ranjay Krishna, Michael Stark, Li-Jia Li, David Shamma, Michael Bernstein, and Li Fei-Fei. Image retrieval using scene graphs. In *CVPR*, 2015. 2

[31] Valentin Khrulkov, Leyla Mirvakhabova, Evgeniya Ustinova, Ivan Oseledets, and Victor Lempitsky. Hyperbolic image embeddings. In *CVPR*, 2020. 8

[32] Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, Li-Jia Li, David A Shamma, et al. Visual genome: Connecting language and vision using crowdsourced dense image annotations. *IJCV*, 2017. 2

[33] Alina Kuznetsova, Hassan Rom, Neil Alldrin, Jasper Uijlings, Ivan Krasin, Jordi Pont-Tuset, Shahab Kamali, Stefan Popov, Matteo Malloci, Alexander Kolesnikov, et al. The open images dataset v4. *IJCV*, 2020. 4, 8

[34] John M Lee. *Introduction to Riemannian manifolds*. Springer, 2018. 3

[35] Tsung-Yi Lin, M. Maire, Serge J. Belongie, James Hays, P. Perona, D. Ramanan, Piotr Dollár, and C. L. Zitnick. Microsoft coco: Common objects in context. In *ECCV*, 2014. 4

[36] Nathan Linial, Eran London, and Yuri Rabinovich. The geometry of graphs and some of its algorithmic applications. *Combinatorica*, 15(2):215–245, 1995. 3

[37] Qi Liu, Maximilian Nickel, and Douwe Kiela. Hyperbolic graph neural networks. *NeurIPS*, 2019. 8

[38] Shaoteng Liu, Jingjing Chen, Liangming Pan, Chong-Wah Ngo, Tat-Seng Chua, and Yu-Gang Jiang. Hyperbolic visual embedding learning for zero-shot recognition. In *CVPR*, 2020. 8

[39] Songtao Liu, Zeming Li, and Jian Sun. Self-emd: Self-supervised object detection without imagenet, 2021. 1, 8

[40] Teng Long, Pascal Mettes, Heng Tao Shen, and Cees G. M. Snoek. Searching for actions on the hyperbole. In *CVPR*, 2020. 8

[41] Thomas Mensink, Efstratios Gavves, and Cees GM Snoek. Costa: Co-occurrence statistics for zero-shot classification. In *CVPR*, 2014. 2

[42] George A Miller, Richard Beckwith, Christiane Fellbaum, Derek Gross, and Katherine J Miller. Introduction to wordnet: An on-line lexical database. *International journal of lexicography*, 3(4):235–244, 1990. 2, 8

[43] Shlok Kumar Mishra, Anshul B. Shah, Ankan Bansal, Jonghyun Choi, Abhinav Shrivastava, Abhishek Sharma, and David Jacobs. Learning visual representations for transfer learning by suppressing texture. *ArXiv*, abs/2011.01901, 2020. 8

[44] Shlok Kumar Mishra, Anshul B. Shah, Ankan Bansal, Abhyuday N. Jagannatha, Abhishek Sharma, David Jacobs, and Dilip Krishnan. Object-aware cropping for self-supervised learning. *ArXiv*, abs/2112.00319, 2021. 4

[45] Maximillian Nickel and Douwe Kiela. Poincaré embeddings for learning hierarchical representations. *NeurIPS*, 2017. 2, 3, 8

[46] Maximillian Nickel and Douwe Kiela. Learning continuous hierarchies in the lorentz model of hyperbolic geometry. In *ICML*, 2018. 8

[47] Curtis Northcutt, Lu Jiang, and Isaac Chuang. Confident learning: Estimating uncertainty in dataset labels. *Journal of Artificial Intelligence Research*, 70:1373–1411, 2021. 6

[48] Devi Parikh and Tsuhan Chen. Hierarchical semantics of objects (hsos). In *ICCV*, 2007. 2, 7

[49] Devi Parikh, C Lawrence Zitnick, and Tsuhan Chen. Unsupervised learning of hierarchical spatial structures in images. In *CVPR*, 2009. 7

[50] Jiwoong Park, Junho Cho, Hyung Jin Chang, and Jin Young Choi. Unsupervised hyperbolic representation learning via message passing auto-encoders. In *CVPR*, 2021. 8

[51] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *ICML*, 2021. 1

[52] Benjamin Recht, Rebecca Roelofs, Ludwig Schmidt, and Vaishaal Shankar. Do imagenet classifiers generalize to imagenet? In *ICML*, 2019. 6

[53] Frederic Sala, Chris De Sa, Albert Gu, and Christopher Ré. Representation tradeoffs for hyperbolic embeddings. In *ICML*, 2018. 2, 3, 8

[54] Ramprasaath R Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *ICCV*, 2017. 8

[55] Vaishaal Shankar, Rebecca Roelofs, Horia Mania, Alex Fang, Benjamin Recht, and Ludwig Schmidt. Evaluating machine accuracy on imagenet. In *ICML*, 2020. 6

[56] Ryohei Shimizu, YUSUKE Mukuta, and Tatsuya Harada. Hyperbolic neural networks++. In *ICLR*, 2021. 8

[57] Dídac Surís, Ruoshi Liu, and Carl Vondrick. Learning the predictability of the future. In *CVPR*, 2021. 8

[58] Yonglong Tian, Dilip Krishnan, and Phillip Isola. Contrastive multiview coding. In *ECCV*, 2020. 5

[59] Alexandru Tifrea, Gary Bécigneul, and Octavian-Eugen Ganea. Poincaré glove: Hyperbolic word embeddings. In *ICLR*. OpenReview, 2018. 2, 3, 8

[60] Dimitris Tsipras, Shibani Santurkar, Logan Engstrom, Andrew Ilyas, and Aleksander Madry. From imagenet to image classification: Contextualizing progress on benchmarks. In *ICML*, 2020. 6

[61] Jasper RR Uijlings, Koen EA Van De Sande, Theo Gevers, and Arnold WM Smeulders. Selective search for object recognition. *IJCV*, 104(2):154–171, 2013. 4

[62] Vijay Vasudevan, Benjamin Caine, Raphael Gontijo-Lopes, Sara Fridovich-Keil, and Rebecca Roelofs. When does dough become a bagel? analyzing the remaining mistakes on imagenet. *arXiv preprint arXiv:2205.04596*, 2022. 6

[63] Wenguan Wang, Tianfei Zhou, Fisher Yu, Jifeng Dai, Ender Konukoglu, and Luc Van Gool. Exploring cross-image pixel contrast for semantic segmentation. In *ICCV*, 2021. 8

[64] Xinlong Wang, Rufeng Zhang, Chunhua Shen, Tao Kong, and Lei Li. Dense contrastive learning for self-supervised visual pre-training. In *CVPR*, 2021. 1, 5, 8

[65] Zhenzhen Weng, Mehmet Giray Ogut, Shai Limonchik, and Serena Yeung. Unsupervised discovery of the long-tail in instance segmentation using hierarchical self-supervision. In *CVPR*, 2021. 8

[66] Yuxin Wu, Alexander Kirillov, Francisco Massa, Wan-Yen Lo, and Ross Girshick. Detectron2. https://github.com/facebookresearch/detectron2, 2019. 5

[67] Zhirong Wu, Yuanjun Xiong, Stella X Yu, and Dahua Lin. Unsupervised feature learning via non-parametric instance discrimination. In *CVPR*, 2018. 1, 2

[68] Jiahao Xie, Xiaohang Zhan, Ziwei Liu, Yew Soon Ong, and Chen Change Loy. Unsupervised object-level representation learning from scene images. In *NeurIPS*, 2021. 1, 4, 5, 8

[69] Zhenda Xie, Yutong Lin, Zheng Zhang, Yue Cao, Stephen Lin, and Han Hu. Propagate yourself: Exploring pixel-level consistency for unsupervised visual representation learning. In *CVPR*, 2021. 1, 8

[70] Jiexi Yan, Lei Luo, Cheng Deng, and Heng Huang. Unsupervised hyperbolic metric learning. In *CVPR*, 2021. 8