Hyperbolic Contrastive Learning for Visual Representations beyond Objects

Songwei Ge\textsuperscript{1}, Shlok Mishra\textsuperscript{1}, Simon Kornblith\textsuperscript{2}, Chun-Liang Li\textsuperscript{2}, David Jacobs\textsuperscript{1,3}

\textsuperscript{1}University of Maryland, College Park, \textsuperscript{2}Google Research, \textsuperscript{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.

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
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”.

6841
Therefore, we propose to study an object-centric hierarchy across the entire dataset. Formally, given a set of images $\mathcal{S} = \{s_1, s_2, \cdots, s_n\}$, $\mathcal{O}_i = \{o_{i1}^1, o_{i1}^2, \cdots, o_{in}^m\}$ are the object bounding boxes contained in the image $s_i$. We define the regions of scene $\mathcal{R}_i = \{r_1^i, r_2^i, \cdots, r_m^i\}$ to be partial areas of the image $s_i$, that contain multiple objects such that $r_1^i = \cup k o_k^i$, where $o_k^i \in \mathcal{O}_i$ and object $k$ is in the region $j$. We define the object-centric hierarchy $T = (V, E)$ to be that $V = \mathcal{S} \cup \dot{\mathcal{O}} \cup \mathcal{R}$, where $\mathcal{R} = \mathcal{R}_1 \cup \cdots \cup \mathcal{R}_n$ and $\dot{\mathcal{O}} = \dot{\mathcal{O}}_1 \cup \cdots \cup \dot{\mathcal{O}}_n$. For $u, v \in V$, $e = (u, v)$ is an edge of $T$ if $u \preceq v$ or $v \preceq u$. 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 $(\mathbb{H}^n, g)$ 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 $\mathbb{D}^n := \{p \in \mathbb{R}^n \mid \|p\|^2 < r^2\}$ 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 $g_{\mathbb{D}}(p) = \frac{4}{(1-\|p\|^2)^2} g_{E}$, where $p \in \mathbb{D}^n$ and $g_{E}$ 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:

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

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 diffeomorphic 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 $\exp_p : T_p \mathbb{H}^n \approx \mathbb{R}^n \rightarrow \mathbb{H}^n$ 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 $\exp_p : T_p \mathbb{D}^n \rightarrow \mathbb{D}^n$ is defined as

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

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 RSVD [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 $p \in \mathbb{H}^n$, there are no conjugate points along any geodesic starting from $p$. 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 $f : T \rightarrow \mathbb{D}^n$ such that for all $i, j$,

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

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. $\|p\| = \|q\|$, the 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 Hyperbolic Contrastive Learning (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 $z_{\text{euc}}^1$ and $z_{\text{euc}}^2$. We follow MoCo [27] and leverage a memory bank to store the negative representations $a_{\text{euc}}^n$, which are the features $z_{\text{euc}}^2$ 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:

$$
L_{\text{euc}} = - \log \frac{\exp \left( z_{\text{euc}}^1 \cdot z_{\text{euc}}^2 / \tau \right)}{\exp \left( z_{\text{euc}}^1 \cdot z_{\text{euc}}^2 / \tau \right) + \sum_n \exp \left( z_{\text{euc}}^1 \cdot a_{\text{euc}}^n / \tau \right)},
$$

where $\tau$ 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 $\mathcal{N}_u = \{v \mid (u, v) \notin E\}$. 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 $z_{\text{hyp}}^1$ and $z_{\text{hyp}}^2$. 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:

$$
\mathcal{L}_{\text{hyp}} = - \log \frac{\exp \left( -d_{\text{hyp}}(z_{\text{hyp}}^1, z_{\text{hyp}}^2) / \tau \right)}{\exp \left( -d_{\text{hyp}}(z_{\text{hyp}}^1, z_{\text{hyp}}^2) / \tau \right) + \sum_n \exp \left( -d_{\text{hyp}}(z_{\text{hyp}}^1, a_{\text{hyp}}^n) / \tau \right)},
$$

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:

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

where $\lambda$ 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](~1.7 million samples) and a subset of OpenImages (~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
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 $L_{hyp}$. This approach, which we denote as “HCL w/o $L_{hyp}$”, 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

![Table 1. Comparison with state-of-the-art methods.](image)

- **MoCo-v2 pre-trained on COCO:**
  - Baseline: 38.5 58.1 42.1 34.8 55.3 37.3
  - HCL w/o $L_{hyp}$: 39.7 60.1 43.4 36.0 57.3 38.8
  - HCL: 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 $L_{hyp}$: 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 $L_{hyp}$: 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 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.

- **MoCo-v2**
  - Pre-train: COCO
  - Bbox: VOC IN-100 IN-1k
  - Results: 64.79 64.84 51.17

- **HCL w/o $L_{hyp}$**
  - Pre-train: COCO
  - Bbox: SS GT
  - Results: 75.55 76.22 54.52

- **HCL**
  - Pre-train: COCO
  - Bbox: GT
  - Results: 76.51 76.74 55.63

**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-training 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.

<table>
<thead>
<tr>
<th>Method</th>
<th>Indicator</th>
<th>Datasets</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>IN-Real</td>
</tr>
<tr>
<td>MoCo</td>
<td>Entropy</td>
<td>0.633</td>
</tr>
<tr>
<td>Supervised</td>
<td>Entropy</td>
<td>0.671</td>
</tr>
<tr>
<td>HCL</td>
<td>Norm</td>
<td>0.655</td>
</tr>
<tr>
<td>Ensemble</td>
<td>Entropy+Norm</td>
<td>0.717</td>
</tr>
</tbody>
</table>

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 a 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 $L_{hyp}$ 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.
Smallest norms (objects) ———— ... ———— Largest norms (scenes)

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 $\lambda$ 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 $\lambda$s 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
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 $\varepsilon$ from $1e^{-5}$ to $1e^{-1}$. This allows a larger $\lambda$ 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. DetCon [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. Belowpar 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


[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.


