
Hyp-OW: Exploiting Hierarchical Structure Learning with Hyperbolic Distance Enhances Open World Object Detection

Thang Doan* Xin Li Sima Behpour Wenbin He Liang Gou Liu Ren
 Bosch Research North America &
 Bosch Center for Artificial Intelligence (BCAI)

Abstract

Open World Object Detection (OWOD) is a challenging and realistic task that extends beyond the scope of standard Object Detection task. It involves detecting both known and unknown objects while integrating learned knowledge for future tasks. However, the level of ‘unknownness’ varies significantly depending on the context. For example, a tree is typically considered part of the background in a self-driving scene, but it may be significant in a household context. We argue that this external or contextual information should already be embedded within the known classes. In other words, there should be a semantic or latent structure relationship between the known and unknown items to be discovered. Motivated by this observation, we propose Hyp-OW, a method that learns and models hierarchical representation of known items through a SuperClass Regularizer. Leveraging this learned representation allows us to effectively detect unknown objects using a Similarity Distance-based Relabeling module. Extensive experiments on benchmark datasets demonstrate the effectiveness of Hyp-OW achieving improvement in both known and unknown detection (up to 6 points). These findings are particularly pronounced in our newly designed benchmark, where a strong hierarchical structure exists between known and unknown objects.

1 Introduction

Advances in Object Detection (OD) have unlocked a plethora of practical applications such as robotics Zhou et al. (2022), self-driving cars Balasubramaniam and Pasricha (2022), manufacturing Malburg et al. (2021), and medical analysis Yang and Yu (2021). Recent breakthroughs in attention-based neural network architecture, such as Deformable Transformers Zhu et al. (2021), have yielded impressive performance in these settings. However, most of these approaches assume a fixed number of classes (closed-world assumption), which is rare in reality. Continual Object Detection Menezes et al. (2023) takes a step further by incrementally adding new classes, resulting in a distribution shift in the input and the well-known phenomenon of *catastrophic forgetting* (Kirkpatrick et al., 2017; Doan et al., 2021) where the network forgets previously learned knowledge. Open World (OW) Bendale and Boult (2015) takes these assumptions even further, introducing the detection and integration of newly discovered classes.

While the seminal work by Bendale and Boult (2015) introduced the Open World (OW) framework, further advancements by Joseph et al. (2021) extended it in two key aspects: the detection task and continual learning. However, a significant challenge within this framework lies in the absence of annotations for unknown objects, leading to biases towards known labels and potential confusion be-

*corresponding author: thang.doan@us.bosch.com

tween unknown items and the background. This bias significantly impedes the accurate identification of unknown objects and presents a major hurdle in the detection process.

We can summarize previous attempts to solve this problem into three main categories. The first category includes works such as Joseph et al. (2021); Gupta et al. (2022); Zohar et al. (2022), which relied on a learned objectness score to relabel the background as potential unknowns. Another direction of research focused on clustering classes to better isolate unknowns (Wu et al., 2022b; Yu et al., 2022). Additionally, some works, like Kim et al. (2022); Wu et al. (2022a), introduced a decoupled approach where the classification and localization heads are separated. The objective is to remove label information and instead capture the shared features that makes these labelled items relevant as objects.

However, these works fail to address a crucial problem, which is defining what constitutes an "unknown" object. Currently, there is no clear definition or prior knowledge available to effectively distinguish unknowns from the background. Its interpretation greatly varies depending on the context. For example, in a driving scene, a "debris on the road" could be considered an unknown object Balasubramaniam and Pasricha (2022), while in a camera surveillance context, it might be perceived as part of the background Ingle and Kim (2022). Without considering the context, these works can only learn to differentiate knowns and unknowns at low level features such as texture or shape. As a consequence, they miss to model any hierarchical structures and similarities between known and unknown items, whether at the image level or dataset level.

Acknowledging this context information, we argue that a hierarchical structural relationship must exist between the objects to be discovered and the known items Hosoya et al. (2022). This hierarchy is characterized by classes that share the same semantic context, belonging to the same category such as vehicles, animals, or electronics. Such hierarchical relationships enable the retrieval of common features and facilitate the discovery of unknown objects. For instance, a model trained on objects related to driving scenes can adequately detect stop signs or traffic lights but is not expected to recognize unrelated objects like a couch or any furniture.

In light of this discrepancy, we propose to learn hierarchical relationships between items in order to effectively utilize the representation of known objects for the discovery of unknown items. Ideally, items belonging to the same family (or category) should be closer to each other while being further away from different families (e.g., animals versus vehicles). To capture these structures, Hyperbolic Distance (Nickel and Kiela, 2018; Park et al., 2021), which naturally maps hierarchical latent structures, such as graphs or trees, emerges as a natural distance metric. This mapping in the hyperbolic space exhibits the desirable property of capturing the affinity between unknown items and known items, thereby enhancing the detection of unknown objects.

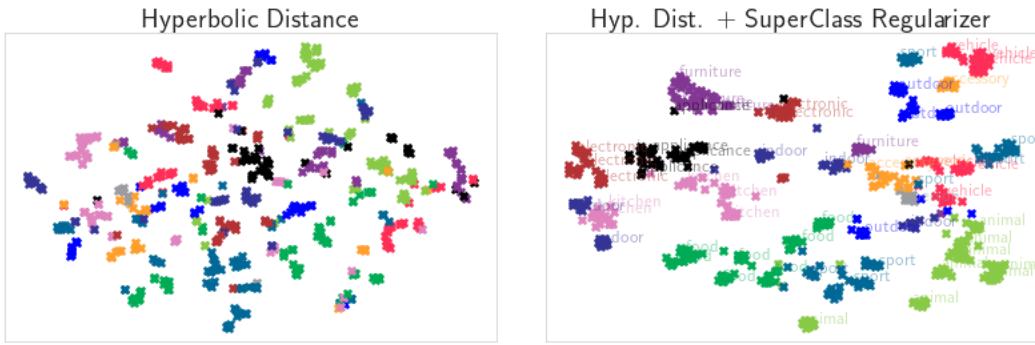


Figure 1: t-SNE plot of the learned class representations, with colors representing their respective categories. Our SuperClass Regularizer (right) enhances the hierarchical structure by grouping together classes from the same category while pushing apart classes from different categories.

Contribution Motivated by the aforementioned literature gap, we propose a **Hyperbolic Distance-based Adaptive Relabeling Scheme** for **Open World** Object Detection (dubbed Hyp-OW). Our contribution can be summarized in three parts:

- Hyp-OW is a simple yet effective method that learns inherent hierarchical structure between objects grouping item from the same category closer while pushing classes from different categories further apart through a SuperClass Regularizer (illustrated in Figure 1, right).
- We propose an Adaptive Relabeling Scheme that enhances the detection of unknown objects by leveraging the semantic similarity between known and unknown objects in the hyperbolic space.
- Our experiments demonstrate significant improvements in both unknown detection (up to 6%) and known accuracy performance (up to 5%) with Hyp-OW . These gains are particularly prominent when evaluating on our (designed) Hierarchical dataset, highlighting the advantages of our method in the presence of high inherent hierarchical structures.

2 Related Work

2.1 Open World Object Detection

The OWOD framework, introduced by Joseph et al. (2021), has inspired many recent works due to its realistic and close-to-real-world setting that integrates newly discovered items into the base knowledge progressively. While the first stream of work was originally based on the Faster-RCNN model (Joseph et al., 2021; Yu et al., 2022; Wu et al., 2022b,a), more recent works have utilized Deformable Transformers due to their superior performance (Gupta et al., 2022; Zohar et al., 2022). Joseph et al. (2021) introduced ORE, a Faster-RCNN-based model that learns class prototypes using contrastive learning with Euclidean distance. However, their approach relied on a held-out validation set where unknown items are explicitly labeled to learn an energy-based model to discriminate unknown items. Yu et al. (2022) extended this setting by minimizing the overlap between the distributions of unknown and known classes. OW-DETR (Gupta et al., 2022) designed a novelty-branch head to relabel the top-k highest background scores as unknowns. These pseudo-labels relied on unmatched bounding box proposals with high backbone activation being selected as unknown objects. On the other hand, Wu et al. (2022a) decoupled the localization and classification tasks (introduced by (Kim et al., 2022)) by learning a class-free head to localize objects. Recently, PROB (Zohar et al., 2022) learned a probabilistic objectness score by learning common statistics for all objects using Mahalanobis distance (Lee et al., 2018) and considered all the remaining bounding box proposals as unknown items. During the evaluation phase, they filter out proposal bounding boxes using the latter probabilistic models.

2.2 Class-Agnostic Object Detection

Another stream of work in the field of object detection is dubbed class-agnostic object detection, which focuses on localizing objects (Kim et al., 2022; Wu et al., 2022a; Jaiswal et al., 2021). The objective is to remove the class label information and learn a shared low-level feature representation that effectively captures the essence of an object. Kim et al. (2022) designed a pure localization head by introducing a second branch that is decoupled from the classification head. Jaiswal et al. (2021) introduced an adversarial objective loss function that penalizes label information in the encoded features. Pixel-wise class-free object detection Gonçalves et al. (2022) used texture gray level quantization to retrieve objects. Saito et al. Saito et al. (2022) designed a new data augmentation method that pastes an annotated object onto an object-free background. Maaz et al. (2022) leveraged language models to improve unknown detection with their Multi-Modal Vision Transformers.

2.3 Learning hierarchical Representation with Hyperbolic Distance

Pointcare embeddings have been widely used in the literature to learn hierarchical structures from complex symbolic or multi-relational data, which can be represented by graphs or trees, such as social networks or taxonomies (Nickel and Kiela, 2018; Law et al., 2019). Due to its good performance, it has been applied to image classification as well (Khrulkov et al., 2020a; Yan et al., 2021; Yue et al., 2023; Ermolov et al., 2022). For example, Yan et al. (2021) used hierarchical clustering to approximate a multi-layered tree structure representation that guides the hyperbolic distance learning process. Similarly, Liu et al. (2020) used taxonomy embedding from GloVe (Pennington et al., 2014) to learn a finer-grained representation. Hyperbolic distance has also been used for object detection (Lang et al., 2022; Ge et al., 2022). Ge et al. (2022) was interested in learning context-object

association rules by reasoning on different image scales. However, none of them leveraged the learned hyperbolic distance to retrieve unknowns items for OWOD.

3 Background

3.1 Problem Formulation

OWOD framework describes the setting where a user receives over the time a stream of T tasks indexed by $t \in [1, T]$. Every task t contains $C_t \in \mathbb{N}^*$ known classes (denoted by set \mathcal{K}^t ²). The goal is to train an object detector module f to accurately recognize the known classes but also discovering unknown classes (denoted by set \mathcal{U}^t). At the end of task t , Δ_t unknown classes are labelled (with an oracle) and included in the next task $t + 1$ (containing now $\sum_{j=1}^t \Delta_j = C_t$ known classes). The process repeat until task T that does not contain anymore unknowns.

The dataset of task t is defined as $\mathcal{D}^t = \{\mathcal{I}^t, \mathcal{Y}^t\}$ where $\mathcal{I}^t = \{\mathbf{I}_{t,1}, \mathbf{I}_{t,2}, \dots, \mathbf{I}_{t,N_t}\}$, are image inputs and $\mathcal{Y}^t = \{\mathbf{y}_{t,1}, \mathbf{y}_{t,2}, \dots, \mathbf{y}_{t,N_t}\}$ are the corresponding labels (There are N_t images for task t). Each image $I_{t,i}$ contains a set of annotations $\mathbf{Y}_{t,i} = [\mathbf{l}_{t,i}, \mathbf{x}_{t,i}, \mathbf{y}_{t,i}, \mathbf{w}_{t,i}, \mathbf{h}_{t,i}]$ where $\mathbf{l}_{t,i} \in \{0, 1\}^C$ denotes the object classes and $[\mathbf{x}_{t,i}, \mathbf{y}_{t,i}, \mathbf{w}_{t,i}, \mathbf{h}_{t,i}]$ are the bounding box coordinates. Throughout the training, we will be storing item in a replay buffer \mathcal{M} with a capacity of m exemplar per class. We denote \mathcal{B} the incoming batch.

We follow the setting of OWOD Joseph et al. (2021) where a set of K exemplars of each class is stored in a replay buffer at the end of each task t (to mitigate forgetting). Those exemplars are then replayed, i.e after task 2, we replay $(2 \cdot K \cdot 20)$ exemplars both coming from task 1 and 2 (K exemplars from each task and 20 classes per task).

3.2 Deformable Transformers for OWOD

We adopt Deformable Transformers (Zhu et al., 2021) as our base detector, as suggested by Gupta et al. (2022), due to its simplicity and high performance. This architecture processes the image input through a set of encoder-decoder modules to produce Q queries output embeddings $\{\mathbf{q}_i\}_{i=1}^Q$, where $\mathbf{q}_i \in \mathbb{R}^d$ (d being the query embedding dimension). These queries are then passed to the bounding box and classification heads, which respectively localize the labeled items and predict their classes. A bipartite matching algorithm (specifically, the Hungarian algorithm (Kuhn, 1955)) is used to match the labeled ground-truth items with each query. The remaining challenge is then to determine whether unmatched queries contain potential unknowns.

3.3 Hyperbolic Embeddings

A Hyperbolic space is a n -dimensional Riemann manifold defined as $(\mathbb{B}_c^n, g^{\mathbb{M}})$ with its Poincare ball $\mathbb{B}_c^n = \{x \in \mathbb{R}^n : c\|x\|^2 \leq 1, c \geq 0\}$ (c being the constant curvature) and equipped with a Riemannian metric $g^{\mathbb{M}} = \lambda_x^2 g^E$ where $\lambda_x^2 = \frac{2}{1-c\|x\|^2}$, $g^E = \mathbf{I}_n$ is the Euclidian metric tensor. The transformation from the Euclidian to hyperbolic space is done via a bijection termed *exponential* mapping $\exp_b^c : \mathbb{R}^n \rightarrow \mathbb{B}_c^n$.

$$\exp_b^c(\mathbf{x}) = \mathbf{b} \otimes_c (\tanh(\sqrt{c} \frac{\lambda_b^c \|\mathbf{x}\|}{2}) \frac{\|\mathbf{x}\|}{\sqrt{c}\|\mathbf{x}\|}) \quad (1)$$

where \mathbf{b} represents the base point. The latter is often empirically taken as $\mathbf{b} = \mathbf{0}$ to simplify the formulas without impacting much the results (Ermolov et al., 2022). We will also adopt this value in our study.

Inside this hyperbolic space, the distance between two points $\mathbf{x}, \mathbf{y} \in \mathbb{B}_c^n$ is computed as:

$$d_{hyp}(\mathbf{x}, \mathbf{y}) = \frac{2}{\sqrt{c}} \arctan(\sqrt{c} \|\mathbf{x} \otimes_c \mathbf{y}\|) \quad (2)$$

where the addition operation \otimes_c is defined as : $\mathbf{x} \otimes_c \mathbf{y} = \frac{(1+2c\langle \mathbf{x}, \mathbf{y} \rangle + c\|\mathbf{y}\|^2)\mathbf{x} + (1-c\|\mathbf{x}\|^2)\mathbf{y}}{1+2c\langle \mathbf{x}, \mathbf{y} \rangle + c^2\|\mathbf{x}\|^2\|\mathbf{y}\|^2}$.

²whenever there is no ambiguity we will remove the task index t to de-clutter the notation

From now on, we will denote \mathbf{z}_i the projection of the queries \mathbf{q}_i into the hyperbolic embedding space, i.e., $\mathbf{z}_i = \exp_{\mathbf{b}}^c(\mathbf{q}_i)$. Note that when $c = 0$, the distance boils down to the cosine similarity defined as:

$$D_{cos}(\mathbf{x}, \mathbf{y}) = 2 - 2 \frac{\langle \mathbf{x}, \mathbf{y} \rangle}{\|\mathbf{x}\|_2 \cdot \|\mathbf{y}\|_2} \quad (\text{in this case no exponential mapping is needed, i.e } \mathbf{z}_i = \mathbf{q}_i^3)$$

4 Hyp-OW

In this section, we provide a detailed explanation of each module of our proposed method. Hyp-OW can be summarized by three main components (Figure 2): a Hyperbolic Metric Distance learning (Section 4.1), a SuperClass Regularizer (Section 4.2) and an Adaptive Relabeling Scheme to detect unknowns (Section 4.3).

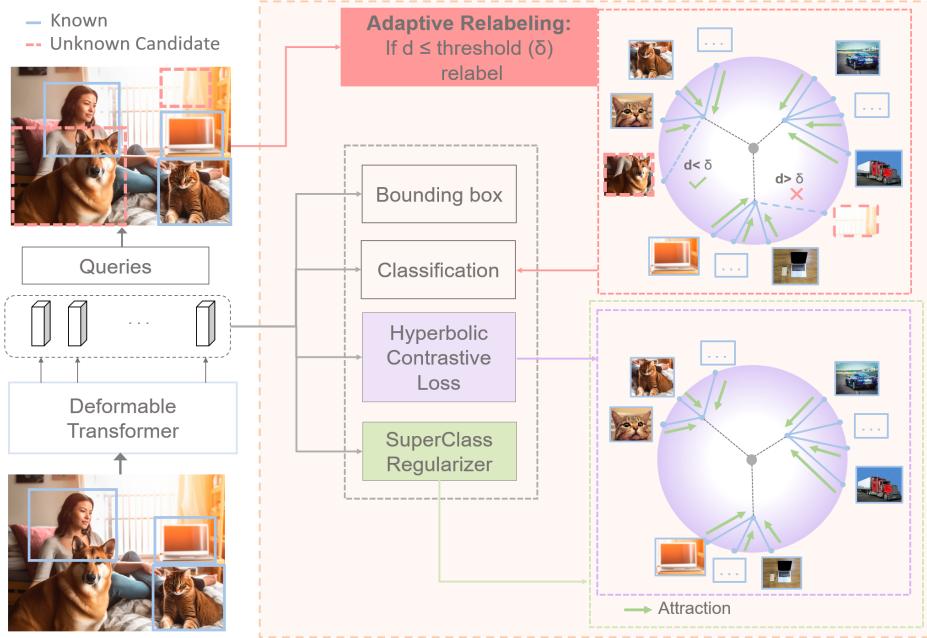


Figure 2: **Overview of each component of Hyp-OW.** There are three main components: the *Hyperbolic Contrastive Loss*, which learns a hierarchical structural representation of each class; the *SuperClass Regularizer*, which models the semantic relationship among classes to ensure proximity within the same category and distance from different categories; and the *Adaptive Relabeling* module, which utilizes the hierarchical structure to compute the hyperbolic distance between candidate proposals and known items. If this distance d is lower than a certain threshold (δ), the proposal is relabelled as unknown.

4.1 Metric Learning with Hyperbolic Distance

We learn feature representation in the hyperbolic embedding space using a contrastive loss. The idea is to move closer features belonging to the same class c while repelling them away from any features from class $j \neq c$. Let's denote \mathbf{z}_i^c any query i matched with class $c \in \mathcal{K}$ ⁴. During the training, we maintain a replay buffer \mathcal{M} where we store m embedding features per class.

For every query element \mathbf{z}_i^c of the incoming batch \mathcal{B} , we sample $k = 1$ elements of the same class $c \in \mathcal{K}$ from the replay buffer \mathcal{M} (which will be considered as the positive examples, the $2|\mathcal{B}| - 2$ remaining samples be considered as the negative examples).

³from now on, we will refer to \mathbf{z}_i interchangeably for both the query and its corresponding embedding \mathbf{q}_i

⁴by abuse of notation, we will also use c for the class label since the meaning can be inferred from the context

For simplicity we denote all the elements from the buffer and the batch $\mathcal{A} = \mathcal{B} \cup \mathcal{M}$. Every element $\mathbf{z}_i^c \in \mathcal{B}$ (respectively $i \in \mathcal{M}$) has its positive counterpart examples \mathbf{z}_{i+} of same class c from \mathcal{M} (respectively from \mathcal{B}) and $|\mathcal{A}| - 2$ negative examples (both from \mathcal{B} and \mathcal{M}) denoted as \mathbf{z}_{i-} .

Defining a temperature τ_1 , the contrastive loss is then expressed as:

$$\mathcal{L}_{hyp} = - \sum_{i \in \mathcal{A}, c \in \mathcal{K}} \log \frac{\exp(\frac{-d_{hyp}(\mathbf{z}_i^c, \mathbf{z}_{i+})}{\tau_1})}{\sum_{i^- \in \mathcal{A} \setminus \{i, i^+\}} \exp(\frac{-d_{hyp}(\mathbf{z}_i^c, \mathbf{z}_{i-})}{\tau_1})} \quad (3)$$

This loss aims at attracting representation of \mathbf{z}_i^c closer to its positive counterpart \mathbf{z}_{i+} while repelling from other classes representation $\mathbf{z}_{i-}, i \in \mathcal{A}, c \in \mathcal{K}$.

4.2 SuperClass Regularization

Many real-world datasets exhibit a natural hierarchical structure, where classes can be organized into categories. For instance, dogs and cats belong to the animal category, while cars and trucks belong to vehicles category. To leverage this hierarchical information, we propose a SuperClass Regularizer (we will use SuperClass and category interchangeably throughout this study), which encourages classes within the same category to be closer in the embedding space while pushing them away from classes in different categories.

Let's denote \mathcal{S}_p the set of class indexes belonging to Category $p = 1 \dots P$. We approximate the category p embedding by computing the Hyperbolic Average (Khrukov et al., 2020b) (dubbed *HypAve*) of every embedding $\{\mathbf{z}_i^c\}_{i \in \mathcal{M}}$ of classes $c \in \mathcal{S}_p$ from the buffer \mathcal{M} , that is:

$$HypAve(\{\mathbf{z}_i^c\}_{i \in \mathcal{M}, c \in \mathcal{S}_p}) = \frac{\sum_{i \in \mathcal{M}, c \in \mathcal{S}_p} \gamma_i \mathbf{z}_i^c}{\sum_{i \in \mathcal{M}} \gamma_i} \quad (4)$$

where $\gamma_i = \frac{1}{\sqrt{1 - c \|\mathbf{x}_i\|^2}}$ are the Lorentz factors. To de-clutter the notation, we will denote $\bar{\mathbf{z}}_p = HypAve(\{\mathbf{z}_i^c\}_{i \in \mathcal{M}, c \in \mathcal{S}_p})$ the Hyperbolic Average feature representation of category p . For every element \mathbf{z}_i^c of a batch \mathcal{B} , we sample its category embedding $\bar{\mathbf{z}}_p (c \in \mathcal{S}_p)$ from the buffer \mathcal{M} and use the same contrastive loss:

$$\mathcal{L}_{reg} = \sum_{i \in \mathcal{A}, c \in \mathcal{S}_p} -\log \frac{\exp(\frac{-d_{hyp}(\mathbf{z}_i^c, \bar{\mathbf{z}}_p)}{\tau_2})}{\sum_{k \neq p} \exp(\frac{-d_{hyp}(\mathbf{z}_i^c, \bar{\mathbf{z}}_k)}{\tau_2})} \quad (5)$$

This loss encourages the features \mathbf{z}_i^c of each class c to be closer to its corresponding category embedding $\bar{\mathbf{z}}_p$, while simultaneously pushing it away from embeddings of other categories $\bar{\mathbf{z}}_k, k \neq p$.

4.3 Adaptive Relabeling of Unknowns with Hyperbolic Distance

We introduce our Adaptive Relabeling module, which *dynamically adapts* to the batch statistics to effectively retrieve unknown items (leveraging the learned hyperbolic embedding). Recall for the classification task, we have two types of queries: matched queries with ground-truth labels (found with the Hungarian Algorithm), and unmatched queries that may contain unknown items. The matched queries will establish a threshold criterion to select unknown queries from the unmatched set.

Let's denote \mathbf{z}^m (respectively \mathbf{z}^u) the query from batch \mathcal{B} that is matched to a ground truth label (respectively that is not matched with any ground truth labels). Next, we denote $\underline{\mathbf{z}}_c$ ⁵ the Hyperbolic Average of class $c \in \mathcal{K}$ computed from samples of the buffer \mathcal{M} as:

⁵we differentiate from $\bar{\mathbf{z}}_p$ with an underline to distinguish Hyperbolic Average of class and category

$$HypAve(\{\mathbf{z}_i^c\}_{i \in \mathcal{M}}) = \frac{\sum_{i \in \mathcal{M}} \gamma_i \mathbf{z}_i^c}{\sum_{i \in \mathcal{M}} \gamma_i}$$

which can be seen as the centroid of each class c in the hyperbolic embedding space.

Next, we define an important quantity: $\delta_{\mathcal{B}} = \max_{m \in \mathcal{B}, c \in \mathcal{K}} d_{hyp}(\mathbf{z}^m, \mathbf{z}_c)$. Intuitively, $\delta_{\mathcal{B}}$ represents the highest distance from every matched query of the batch \mathcal{B} to all centroid $\mathbf{z}_c, c \in \mathcal{K}$ from the replay buffer \mathcal{M} . The latter will serve as a threshold to relabel every unmatched query \mathbf{z}^u as unknown if:

$$\min_{c \in \mathcal{K}} d_{hyp}(\mathbf{z}^u, \mathbf{z}_c) \leq \delta_{\mathcal{B}} \quad (6)$$

The underlying idea is that if any unmatched query \mathbf{z}^u has a distance to any centroid smaller than $\delta_{\mathcal{B}}$, it is likely to be an unknown. It will then be relabeled accordingly and forwarded to the classification head.

Overall loss All the aforementioned losses are finally optimized together as:

$$\mathcal{L} = \mathcal{L}_{cls} + \mathcal{L}_{bbox} + \alpha \mathcal{L}_{hyp} + \beta \mathcal{L}_{reg} \quad (7)$$

Where $\alpha, \beta \geq 0$ are coefficient controlling respectively the Hyperbolic and Regularizer importance.

5 Experiments

In this section, we begin by describing our experimental setup. We then present comparative results against benchmark baselines, followed by in-depth ablation analysis of each component of Hyp-OW . Due to space limitations, we will defer detailed information to the Appendix Section C and D.

5.1 Experimental Setup

Datasets We consider two benchmarks from the literature: the OWOD Split Joseph et al. (2021) and the OWDETR Split Gupta et al. (2022). While the latter (OWDETR Split) strictly separates superclasses across tasks the first (OWOD) has mild semantic overlap between knowns and unknowns across tasks (See Appendix Section A). To validate our hypothesis regarding the semantic relationship between knowns and unknowns, we introduce a third dataset called the *Hierarchical Split*. This dataset ensures that each task includes at least one class from each category, promoting a higher level of semantic similarity. This will be discussed in paragraph 5.2. Each dataset is defined by four tasks $t = 1, 2, 3, 4$, containing 20 labelled classes each, for a total of 80 classes. When task t starts, only the label of classes belonging to that task are revealed. For instance, task 1 only contains labels of classes from 0 to 19, while task 2 only contains labels of classes from 21 to 39, and so on.

Implementation Details We use Deformable DETR (Zhu et al., 2021) pretrained in a self-supervised manner (DINO Caron et al. (2021)) on Resnet-50 (He et al., 2016) as our backbone. The number of deformable transformer encoder and decoder layers are set to 6. The number of queries is set to $Q = 100$ with a dimension $d = 256$. During inference time, the top-100 high scoring queries per image are used for evaluation. Additional detail are provided in Appendix Section B.

Metrics and Baselines Following current metrics used for OWOD, we use mean average precision (mAP) for known items while U-Recall is the main metric used to quantify the unknown detection quality or each method Gupta et al. (2022); Wu et al. (2022a); Zohar et al. (2022); Maaz et al. (2022); Yu et al. (2022). Additional metric is discussed in Table 9 We consider the following baselines from literature: ORE-EBUI (Joseph et al., 2021), UC-OWOD (Wu et al., 2022b), OCPL (Yu et al., 2022), 2B-OCD (Wu et al., 2022a), OW-DETR (Gupta et al., 2022) and PROB (Zohar et al., 2022).

5.2 Benchmark Results

Dataset Similarity To gain insights into the structure of each dataset, we introduce a semantic similarity measure based on GloVe’s embedding Pennington et al. (2014) (defined in Appendix Section A.1). This measure quantifies the similarity overlap between known and unknown items,

with higher values indicating larger overlap. The three splits, OW-DETR Split (**Low regime**), OWOD Split (**Medium regime**), and Hierarchical Split (**High regime**), exhibit a monotonic ranking in terms of similarity, with respective values of 0.27, 0.33, and 0.41. This serves as a starting point for evaluating baseline methods under various scenarios.

Unknown Detection (U-Recall)

Table 1 shows the high performance gain of Hyp-OW over PROB on **Medium regime** and **High regime** of 3% on average. This highlights the utility of learning hierarchical structured representations and retrieving unknowns based on their similarity with known objects, as opposed to PROB, which learns a single mean representation for all objects. For the **Low regime** our method is performing on-par with PROB except for task 1 which shows a surprising improvement of 6 points. A possible explanation may come from the nature of the Object Detection (OD) task, where there can be significant overlap between bounding boxes. This encourages the model to learn classes that frequently co-occur, such as "person" and "backpack" or "teddy bear" and "bicycle" (See Figure 3). We provide qualitative and quantitative explanations in Appendix Section C.

		Task 1		Task 2		Task 3		Task 4	
		U-Recall (\uparrow)	mAP (\uparrow)	U-Recall (\uparrow)	mAP (\uparrow)	U-Recall (\uparrow)	mAP (\uparrow)	U-Recall (\uparrow)	mAP (\uparrow)
Low	ORE - EBU1	1.5	61.4	3.9	40.6	3.6	33.7	31.8	
	OW-DETR	5.7	71.5	6.2	43.8	6.9	38.5	33.1	
	PROB	17.6	73.4 (+0.7)	22.3	50.4	24.8	42.0	39.9	
	Hyp-OW (Ours)	23.9 (+6.3)	72.7	23.3 (+1.0)	50.6	25.4	46.2 (+4.2)	44.8 (+4.9)	
Medium	ORE - EBU1	4.9	56.0	2.9	39.4	3.9	29.7	25.3	
	UC-OWOD	2.4	50.7	3.4	8.7	16.3	24.6	23.2	
	OCPL	8.26	56.6	7.65	39.1	11.9	30.7	26.7	
	2B-OCD	12.1	56.4	9.4	38.5	11.6	29.2	25.8	
	OW-DETR	7.5	59.2	6.2	42.9	5.7	30.8	27.8	
	PROB	19.4	59.5	17.4	44.0	19.6	36.0	31.5	
High	Hyp-OW (Ours)	23.5 (+4.1)	59.4	20.6 (+3.2)	44.4	26.3 (+6.7)	36.8	33.6 (+2.1)	
	OW-DETR	7.0	47.3	11.0	38.6	8.8	38.3	38.2	
	PROB	29.4	49.6	43.9	42.9	52.7	41.3	41.0	
	Hyp-OW (Ours)	34.9 (+5.5)	49.9	47.5 (+3.6)	45.5 (+2.6)	55.2 (+2.5)	44.3 (+3.1)	43.9 (+2.9)	

Table 1: **State-of-the-art comparison on the three splits for unknown detection (U-Recall) and known accuracy (mAP).** Hyp-OW improves significantly the unknown detection (U-Recall) for the medium and high regime and known detection (mAP) for the low regime. Task 4 does not have U-Recall since all 80 classes are known at this stage.

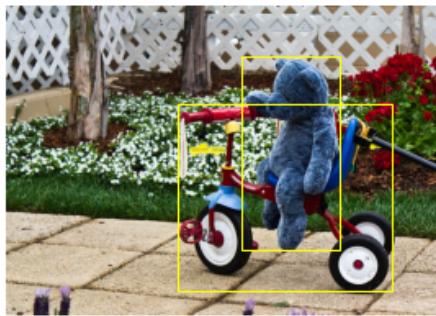


Figure 3: **Bounding boxes overlap.** In the OD task, the high overlap between bounding boxes of frequently co-occurring objects can influence the model to learn their associations and correlations.

Known Accuracy (mAP)

From the known accuracy, Hyp-OW surpasses baseline benchmark on all tasks of Hierarchical Split and shows notable performance for the two last tasks of OW-DETR Split. This can be credited to the structural hierarchy learned that groups together class of same category (See t-SNE Figure 1 middle and right).

5.3 Ablation Analysis

We now propose an in-depth understanding of Hyp-OW by removing one by one each component and see its direct impact (illustrated quantitatively in Table 2 for Hierarchical Split).

	Task 1		Task 2		Task 3		Task 4
	U-Recall (\uparrow)	mAP (\uparrow)	U-Recall (\uparrow)	mAP (\uparrow)	U-Recall (\uparrow)	mAP (\uparrow)	mAP (\uparrow)
Hyp-OW (Ours)	34.9	49.9	47.5	45.5	55.2	44.3	43.9
w/ Cosine Distance	32.8 (-2.1)	49.0 (-0.9)	46.4 (-1.1)	45.4	55.4	43.2 (-1.1)	43.1
w/o SuperClass Regularizer	32.0 (-2.9)	50.0	47.1	45.1	52.9 (-2.3)	43.7	43.5
w/o Adaptive Relabeling	34.7	41.2 (-8.7)	47.6	38.9 (-6.6)	54.1 (-1.1)	36.5 (-7.8)	36.1 (-7.8)

Table 2: **Impact of each component of Hyp-OW on hierarchical Split.** We observe that the Relabeling module (third line) significantly reduces mAP while maintaining U-Recall. On the other hand, the SuperClass Regularizer and Cosine Distance have primarily an impact on unknown detection.

Adaptive Relabeling: This module uses Eq 6 to relabel unmatched bounding boxes as unknowns. To assess the impact of this relabeling strategy, we compare it with an alternative approach used by PROB Zohar et al. (2022), where all unmatched queries are labeled as unknowns. Although the decrease in U-Recall is marginal, we observe a significant reduction in known accuracy (mAP). This performance degradation can be attributed to the over-prediction of patches as unknowns, which results in misclassification of known objects. Heatmaps in Figure 10 and 11 (Appendix) demonstrate the effectiveness of our module where we observe that unknowns belonging to the same category (sharing some color) as knowns exhibit lower Hyperbolic Distance.

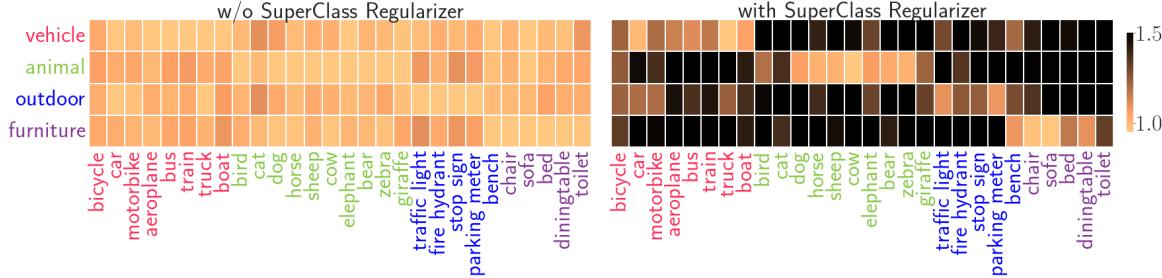


Figure 4: **Hyperbolic Category - Class Distance Heatmap.** The SuperClass Regularizer (right) effectively separates different categories (left), as indicated by the increased distance between each animal class (bottom) and the vehicle, outdoor, and furniture categories (darker colors). Without this Regularizer (left), category inter-distance are much smaller (lighter color intensity).

SuperClass Regularizer: By setting $\beta = 0$, we no longer enforce the grouping of items from the same category in the hyperbolic space (compare t-SNE plots in Figure 1). As a result, we observe a reduction in U-Recall of 2.9, 0.4, and 2.3 points, respectively (Table 2, second line). Heatmap Figure 4 illustrates the hyperbolic distance from each class to every category’s embedding (computed using Eq 4) with lighter colors indicating smaller distances. With our Regularizer (right plot), we observe a wider range of values spanning from 0.7 to 2.30, compared to a smaller range of 0.78 to 1.2 without the Regularizer. This highlights the effect of our Regularizer that pushes classes from different categories apart (shown as dark color in the right plot) while bringing classes of similar categories closer together. A more detailed plot can be found in Appendix Figure 12.

Cosine Distance: Trading hyperbolic distance for the cosine distance ($c = 0$) harms both the U-Recall and mAP as the hyperbolic embedding space is more suitable to learn data with latent hierarchical structure (highlighted by Figure 14 in Appendix).

6 Conclusion

The Open World Object Detection framework presents a challenging and promising setting, encompassing crucial aspects such as lifelong learning and unknown detection. In our work, we have emphasized the lack of a clear definition of unknowns and the need of a hierarchical or semantic relationship between known and unknown classes. This led us to propose Hyp-OW that focuses on learning and modeling the structural hierarchy within the dataset, which is then utilized for unknowns retrieval. Extensive experiments demonstrates significant improvement of Hyp-OW for both known and unknown detection (up to 6 points) particularly in the presence of inherent hierarchy between classes.

Acknowledgement

We would like to thank (in no particular order): Jiajing Guo, Shabnam Ghaffarzadegan, Yulian Guo, Sharath Gopal, Sanbao Bu and Jorge Ono for useful discussions and feedbacks throughout the work.

References

- Abhishek Balasubramaniam and Sudeep Pasricha. Object detection in autonomous vehicles: Status and open challenges, 2022.
- Abhijit Bendale and Terrance Boult. Towards open world recognition. *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Jun 2015. doi: 10.1109/cvpr.2015.7298799. URL <http://dx.doi.org/10.1109/CVPR.2015.7298799>.
- Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 9650–9660, 2021.
- Thang Doan, Mehdi Abbana Bennani, Bogdan Mazoure, Guillaume Rabusseau, and Pierre Alquier. A theoretical analysis of catastrophic forgetting through the ntk overlap matrix. In Arindam Banerjee and Kenji Fukumizu, editors, *Proceedings of The 24th International Conference on Artificial Intelligence and Statistics*, volume 130 of *Proceedings of Machine Learning Research*, pages 1072–1080. PMLR, 13–15 Apr 2021. URL <https://proceedings.mlr.press/v130/doan21a.html>.
- 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.
- Mark Everingham, Luc Van Gool, Christopher KI Williams, John Winn, and Andrew Zisserman. The pascal visual object classes (voc) challenge. *International journal of computer vision*, 88:303–338, 2010.
- Songwei Ge, Shlok Mishra, Simon Kornblith, Chun-Liang Li, and David Jacobs. Hyperbolic contrastive learning for visual representations beyond objects. *arXiv preprint arXiv:2212.00653*, 2022.
- Gabriel R. Gonçalves, Jessica Sena, William Robson Schwartz, and Carlos Antonio Caetano. Pixel-level class-agnostic object detection using texture quantization. In *2022 35th SIBGRAPI Conference on Graphics, Patterns and Images (SIBGRAPI)*, volume 1, pages 31–36, 2022. doi: 10.1109/SIBGRAPI5357.2022.9991762.
- Akshita Gupta, Sanath Narayan, KJ Joseph, Salman Khan, Fahad Shahbaz Khan, and Mubarak Shah. Ow-detr: Open-world detection transformer. In *CVPR*, 2022.
- 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.

Yusuke Hosoya, Masanori Suganuma, and Takayuki Okatani. More practical scenario of open-set object detection: Open at category level and closed at super-category level. *ArXiv*, abs/2207.09775, 2022.

Palash Yuvraj Ingle and Young-Gab Kim. Real-time abnormal object detection for video surveillance in smart cities. *Sensors*, 22(10), 2022. ISSN 1424-8220. doi: 10.3390/s22103862. URL <https://www.mdpi.com/1424-8220/22/10/3862>.

Ayush Jaiswal, Yue Wu, Pradeep Natarajan, and Premkumar Natarajan. Class-agnostic object detection. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 919–928, 2021.

K J Joseph, Salman Khan, Fahad Shahbaz Khan, and Vineeth N Balasubramanian. Towards open world object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR 2021)*, 2021.

Valentin Khrulkov, Leyla Mirvakhabova, Evgeniya Ustinova, Ivan Oseledets, and Victor Lempitsky. Hyperbolic image embeddings. In *The IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020a.

Valentin Khrulkov, Leyla Mirvakhabova, Evgeniya Ustinova, Ivan Oseledets, and Victor Lempitsky. Hyperbolic image embeddings. In *The IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020b.

Dahun Kim, Tsung-Yi Lin, Anelia Angelova, In So Kweon, and Weicheng Kuo. Learning open-world object proposals without learning to classify. *IEEE Robotics and Automation Letters (RA-L)*, 2022.

James Kirkpatrick, Razvan Pascanu, Neil Rabinowitz, Joel Veness, Guillaume Desjardins, Andrei A. Rusu, Kieran Milan, John Quan, Tiago Ramalho, Agnieszka Grabska-Barwinska, Demis Hassabis, Claudia Clopath, Dharshan Kumaran, and Raia Hadsell. Overcoming catastrophic forgetting in neural networks. *Proceedings of the National Academy of Sciences*, 114(13):3521–3526, 2017. doi: 10.1073/pnas.1611835114. URL <https://www.pnas.org/doi/abs/10.1073/pnas.1611835114>.

Harold W. Kuhn. The Hungarian Method for the Assignment Problem. *Naval Research Logistics Quarterly*, 2(1–2):83–97, March 1955. doi: 10.1002/nav.3800020109.

Christopher Lang, Alexander Braun, Lars Schillingmann, and Abhinav Valada. On hyperbolic embeddings in object detection. In *Pattern Recognition: 44th DAGM German Conference, DAGM GCPR 2022, Konstanz, Germany, September 27–30, 2022, Proceedings*, pages 462–476. Springer, 2022.

Marc Law, Renjie Liao, Jake Snell, and Richard Zemel. Lorentzian distance learning for hyperbolic representations. In Kamalika Chaudhuri and Ruslan Salakhutdinov, editors, *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pages 3672–3681. PMLR, 09–15 Jun 2019. URL <https://proceedings.mlr.press/v97/law19a.html>.

Kimin Lee, Kibok Lee, Honglak Lee, and Jinwoo Shin. A simple unified framework for detecting out-of-distribution samples and adversarial attacks. *Advances in neural information processing systems*, 31, 2018.

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. *Lecture Notes in Computer Science*, page 740–755, 2014. ISSN 1611-3349. doi: 10.1007/978-3-319-10602-1_48. URL http://dx.doi.org/10.1007/978-3-319-10602-1_48.

Shaoteng Liu, Jingjing Chen, Liangming Pan, Chong-Wah Ngo, Tat-Seng Chua, and Yu-Gang Jiang. Hyperbolic visual embedding learning for zero-shot recognition. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9273–9281, 2020.

Muhammad Maaz, Hanoona Rasheed, Salman Khan, Fahad Shahbaz Khan, Rao Muhammad Anwer, and Ming-Hsuan Yang. Class-agnostic object detection with multi-modal transformer. In *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part X*, pages 512–531. Springer, 2022.

Lukas Malburg, Manfred-Peter Rieder, Ronny Seiger, Patrick Klein, and Ralph Bergmann. Object detection for smart factory processes by machine learning. *Procedia Computer Science*, 184: 581–588, 2021. ISSN 1877-0509. doi: <https://doi.org/10.1016/j.procs.2021.04.009>. URL <https://www.sciencedirect.com/science/article/pii/S1877050921007821>. The 12th International Conference on Ambient Systems, Networks and Technologies (ANT) / The 4th International Conference on Emerging Data and Industry 4.0 (EDI40) / Affiliated Workshops.

Angelo G Menezes, Gustavo de Moura, Cézanne Alves, and André CPLF de Carvalho. Continual object detection: a review of definitions, strategies, and challenges. *Neural Networks*, 2023.

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.

Jiwoong Park, Junho Cho, Hyung Jin Chang, and Jin Young Choi. Unsupervised hyperbolic representation learning via message passing auto-encoders. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5516–5526, 2021.

Jeffrey Pennington, Richard Socher, and Christopher Manning. GloVe: Global vectors for word representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543, Doha, Qatar, October 2014. Association for Computational Linguistics. doi: 10.3115/v1/D14-1162. URL <https://aclanthology.org/D14-1162>.

Kuniaki Saito, Ping Hu, Trevor Darrell, and Kate Saenko. Learning to detect every thing in an open world. In *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XXIV*, pages 268–284. Springer, 2022.

Yan Wu, Xiaowei Zhao, Yuqing Ma, Duorui Wang, and Xianglong Liu. Two-branch objectness-centric open world detection. In *Proceedings of the 3rd International Workshop on Human-Centric Multimedia Analysis, HCMA ’22*, page 35–40, New York, NY, USA, 2022a. Association for Computing Machinery. ISBN 9781450394925. doi: 10.1145/3552458.3556453. URL <https://doi.org/10.1145/3552458.3556453>.

Zhiheng Wu, Yue Lu, Xingyu Chen, Zhengxing Wu, Liwen Kang, and Junzhi Yu. Uc-owod: Unknown-classified open world object detection. In *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part X*, pages 193–210. Springer, 2022b.

Jiexi Yan, Lei Luo, Cheng Deng, and Heng Huang. Unsupervised hyperbolic metric learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12465–12474, 2021.

Ruixin Yang and Yingyan Yu. Artificial convolutional neural network in object detection and semantic segmentation for medical imaging analysis. *Frontiers in Oncology*, 11, 2021. ISSN 2234-943X. doi: 10.3389/fonc.2021.638182. URL <https://www.frontiersin.org/articles/10.3389/fonc.2021.638182>.

Jinan Yu, Liyan Ma, Zhenglin Li, Yan Peng, and Shaorong Xie. Open-world object detection via discriminative class prototype learning. In *2022 IEEE International Conference on Image Processing (ICIP)*, pages 626–630, 2022. doi: 10.1109/ICIP46576.2022.9897461.

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

Zhengxue Zhou, Leihui Li, Alexander Fürsterling, Hjalte Joshua Durocher, Jesper Mouridsen, and Xuping Zhang. Learning-based object detection and localization for a mobile robot manipulator in sme production. *Robotics and Computer-Integrated Manufacturing*, 73:102229, 2022. ISSN 0736-5845. doi: <https://doi.org/10.1016/j.rcim.2021.102229>. URL <https://www.sciencedirect.com/science/article/pii/S0736584521001113>.

Xizhou Zhu, Weijie Su, Lewei Lu, Bin Li, Xiaogang Wang, and Jifeng Dai. Deformable {detr}: Deformable transformers for end-to-end object detection. In *International Conference on Learning Representations*, 2021. URL <https://openreview.net/forum?id=gZ9hCDWe6ke>.

Orr Zohar, Kuan-Chieh Wang, and Serena Yeung. Prob: Probabilistic objectness for open world object detection, 2022.

Appendix

The Appendix Section is organized as follow:

- Section A provides information on the composition of each dataset and the introduction of our semantic similarity measure.
- Section B presents the experimental details.
- Section C includes comprehensive benchmark metrics results and visual plots (heatmaps) to illustrate our findings.
- Section D discusses the results of the ablation analysis accompanied by visual figures for better understanding.

A Dataset Information

A.1 Dataset Composition

In this section, we provide the composition of each dataset split and explain the main differences between them. The splits are made with images taken from the PASCAL-VOC2007/2012 (Everingham et al., 2010) and MS-COCO dataset (Lin et al., 2014), containing a total of 80 classes grouped into 12 categories (SuperClass): **vehicle**, **outdoor**, **animal**, **accessory**, **sports**, **kitchen**, **food**, **furniture**, **appliance**, **electronic**, **indoor**, and **person**. We categorize the different datasets into three regimes based on a semantic similarity metric (defined below), which quantifies the overlap between known and unknown items:

- **Low regime** : OWDETR Split
- **Medium regime** : OWOD Split
- **High regime** : Hierarchical Split

The low regime implies no semantic similarity between classes of different tasks. For instance, OWDETR Split’s (3) classes of task 1 are composed of vehicle, animal, and person, which do not appear in tasks 2, 3, or 4. On the other hand, high regime shows high similarity between classes across all tasks. Each task of the Hierarchical split (Table 5) contains at least one class from each of the 12 categories. The OWOD Split is a trade-off between the aforementioned datasets.

Low regime : OWDETR Split			
Task 1	Task 2	Task 3	Task 4
aeroplane	traffic light	frisbee	laptop
bicycle	fire hydrant	skis	mouse
bird	stop sign	snowboard	remote
boat	parking meter	sports ball	keyboard
bus	bench	kite	cell phone
car	chair	baseball bat	book
cat	diningtable	baseball glove	clock
cow	pottedplant	skateboard	vase
dog	backpack	surfboard	scissors
horse	umbrella	tennis racket	teddy bear
motorbike	handbag	banana	hair drier
sheep	tie	apple	toothbrush
train	suitcase	sandwhich	wine glass
elephant	microwave	orange	cup
bear	oven	broccoli	fork
zebra	toaster	carrot	knife
giraffe	sink	hot dog	spoon
truck	refrigerator	pizza	bowl
person	bed	donut	tvmonitor
	toilet	cake	bottle
	sofa		

Table 3: **Composition of OWDETR Split.** There is no overlap of categories between each task

Medium regime : OWOD Split			
Task 1	Task 2	Task 3	Task 4
aeroplane	truck	frisbee	bed
bicycle	traffic light	skis	toilet
bird	fire hydrant	snowboard	laptop
boat	stop sign	sports ball	mouse
bottle	parking meter	kite	remote
bus	bench	baseball bat	keyboard
car	elephant	baseball glove	cell phone
cat	bear	skateboard	book
chair	zebra	surfboard	clock
cow	giraffe	tennis racket	vase
diningtable	backpack	banana	scissors
dog	umbrella	apple	teddy bear
horse	handbag	sandwhich	hair drier
motorbike	tie	hot dog	toothbrush
person	suitcase	broccoli	wine glass
pottedplant	microwave	carrot	cup
sheep	oven	hot dog	fork
sofa	toaster	pizza	knife
train	sink	donut	spoon
tvmonitor	refrigerator	cake	bowl

Table 4: **Composition of OWOD Split.** There is a mild overlap of categories between each task.

High regime : hierarchical Split			
Task 1	Task 2	Task 3	Task 4
bicycle	person	bus	truck
car	motorbike	train	boat
traffic light	aeroplane	parking meter	bench
fire hydrant	stop sign	cow	zebra
bird	horse	elephant	giraffe
cat	sheep	bear	tie
dog	umbrella	handbag	suitcase
backpack	snowboard	kite	skateboard
frisbee	sports ball	baseball bat	surfboard
skis	cup	baseball glove	tennis racket
bottle	sandwich	fork	spoon
wine glass	orange	knife	bowl
banana	broccoli	carrot	pizza
apple	pottedplant	hot dog	donut
chair	bed	diningtable	cake
sofa	laptop	remote	toilet
tvmonitor	mouse	keyboard	cell phone
microwave	toaster	sink	refrigerator
oven	clock	scissors	hair drier
book	vase	teddy bear	toothbrush

Table 5: **Composition of hierarchical Split.** Each task contains at least one class of each category.

Quantifying Semantic Overlap between Knowns and Unknowns across Dataset We propose a measure to quantify the semantic similarity overlap between known and unknown classes in each dataset. To achieve this, we utilize the GloVe (Pennington et al., 2014) embedding for each known class $c \in \mathcal{K}^t$ (and unknown class $k \in \mathcal{U}^t$), denoted by ω_c (respectively ω_k). For a given task $t \in [1, T]$, we define the semantic overlap between knowns and unknowns as:

$$S_t = \frac{1}{|\mathcal{U}^t|} \sum_{k \in \mathcal{U}^t} \max_{c \in \mathcal{K}^t} \frac{\langle \omega_c, \omega_k \rangle}{\|\omega_c\|_2 \cdot \|\omega_k\|_2}, \forall t \leq T - 1 \quad (8)$$

Higher value means higher similarity between the knowns and unknowns. For a given task t , this value calculates the similarity between each unknown and the highest-ranked known and averages it across all unknowns. Figure 5 shows the evolution of this similarity measure throughout the training. As expected, the three splits ranks monotonically as low similarity (OWDETR Split), medium similarity (OWOD Split) and high similarity (Our hierarchical Split).

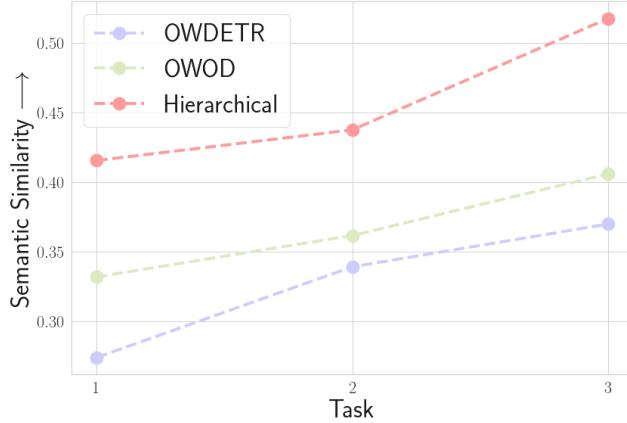


Figure 5: Semantic Similarity between knowns and unknowns across tasks for each Split.

B Implementation Details

In this section, we provide a recap of the definition and value of each hyperparameter used, as well as the learning rate schedule employed during training. The sequence of learning rate for each task is recalled in Table 7. Additional details are provided regarding the usage of the buffer in various loss functions.

Hyperparameters		
Parameters	Value	Definition
α	0.05	Coefficient of the Hyperbolic Contrastive Loss Eq 3
β	0.02	Coefficient of the SuperClass Regularizer Eq 5
τ_1	0.2	Temperature of the Hyperbolic Contrastive Loss
τ_2	0.4	Temperature of the SuperClass Regularizer
m	10	Capacity of the replay buffer (examplar per class)
c	0.1	curvature coefficient for the Hyperbolic Distance
k	1	number of positive examples to be sampled for each anchor
batch size	3	N/A
GPUs	4 Nvidia RTX 3090	N/A

Table 6: **Hyperparameters used for Hyp-OW .**

hierarchical Split			
Task \mathcal{T}_t	From Epoch to	Learning rate l_r	Classes labeled
Task 1	0 to 40	10^{-4}	0-19
Task 1	40 to 50	10^{-5}	0-19
Task 2	50 to 70	10^{-4}	20-39
Task 2 fine-tuning	70 to 130	10^{-4}	0-39
Task 3	130 to 150	10^{-4}	40-59
Task 3 fine-tuning	150 to 210	10^{-4}	0-59
Task 4	210 to 230	10^{-4}	60-79
Task 4 fine-tuning	230-300	10^{-4}	0-80 (no unknowns)

Table 7: **Learning rate schedule throughout the 4 tasks.**

Buffer Filling: Throughout the training process, we store the embeddings of every encountered class \mathbf{z}_i^c (where $c \in \mathcal{K}, i \in \mathcal{B}$) which has a capacity of m exemplars per class.

Buffer Interaction with Hyperbolic Contrastive Loss: To compute the contrastive loss (Eq. 3), for each embedding z_i^c in the batch, where $i \in \mathcal{B}$ and $c \in \mathcal{K}$, we sample its positive counterpart \mathbf{z}_{i+} from the same class c in the buffer.

Buffer Interaction with SuperClass Regularizer: To compute the Hyperbolic Average $\bar{\mathbf{z}}_p$ ($p = 1 \dots P$) in the SuperClass Regularizer (Eq. 5), we sample from the buffer every embedding \mathbf{z}_i^c , $i \in M, c \in \mathcal{S}_p$ then use Eq 4 to calculate $\bar{\mathbf{z}}_p$.

C Additional Experimental Results

C.1 Detailed Results

Table 8 presents a detailed overview of the performance of all baselines across the three splits. It includes the mean Average Precision (mAP) of the Previous and Current known classes for tasks 2, 3, and 4. Hyp-OW exhibits a high level of plasticity, as evidenced by its accuracy on the current known classes (highlighted in the 'Current' column of tasks 2, 3, and 4). This can be attributed to the hierarchical structure learned within the hyperbolic embedding space, which facilitates effective modeling and representation of the known classes.

Regime		Task 1			Task 2			Task 3			Task 4			
		U-Recall	mAP (\uparrow) Current	U-Recall (\uparrow)	Previous	mAP (\uparrow) Current	Both	U-Recall (\uparrow)	Previous	mAP (\uparrow) Current	Both	Previous	mAP (\uparrow) Current	Both
Low	ORE - EBUI	1.5	61.4	3.9	56.5	26.1	40.6	3.6	38.7	23.7	33.7	33.6	26.3	31.8
	OW-DETR	5.7	71.5	6.2	62.8	27.5	43.8	6.9	45.2	24.9	38.5	38.2	28.1	33.1
	PROB	17.6	73.4	22.3	66.3	36.0	50.4	24.8	47.8	30.4	42.0	42.6	31.7	39.9
	Hyp-OW (Ours)	23.9 (+6.3)	72.7	23.3 (+1.0)	59.8	42.2	50.6	25.4	49.3	39.8	46.2 (+4.2)	46.4	40.3	44.8 (+4.9)
Medium	ORE - EBUI	4.9	56.0	2.9	52.7	26.0	39.4	3.9	38.2	12.7	29.7	29.6	12.4	25.3
	UC-OWOD	2.4	50.7	3.4	33.1	30.5	31.8	8.7	28.8	16.3	24.6	25.6	15.9	23.2
	OCPL	8.26	56.6	7.65	50.6	27.5	39.1	11.9	38.7	14.7	30.7	30.7	14.4	26.7
	2B-OCD	12.1	56.4	9.4	51.6	25.3	38.5	11.6	37.2	13.2	29.2	30.0	13.3	25.8
High	OW-DETR	7.5	59.2	6.2	53.6	33.5	42.9	5.7	38.3	15.8	30.8	31.4	17.1	27.8
	PROB	19.4	59.5	17.4	55.7	32.2	44.0	19.6	43.0	22.2	36.0	35.7	18.9	31.5
	Hyp-OW (Ours)	23.5 (+4.1)	59.4	20.6 (+3.1)	51.4	37.4	44.4	26.3 (+6.6)	42.9	24.6	36.8	37.4	22.4	33.6 (+2.1)
	OW-DETR	7.0	47.3	11.0	38.0	39.2	38.6	8.8	39.3	36.1	38.3	38.5	37.3	38.2
	PROB	29.4	49.6	43.9	41.9	43.9	42.9	52.7	41.7	40.4	41.3	40.9	41.2	41.0
	Hyp-OW (Ours)	34.9 (+5.2)	49.9	47.5 (+3.5)	42.0	49.0	45.5 (+2.7)	55.2 (+2.5)	44.4	44.1	44.3 (+3.1)	42.8	47.0	43.9 (+3.0)

Table 8: Benchmark Results.

Next, we turn our attention to the issue of unknown object confusion and address it through the Absolute Open-Set Error (A-OSE) metric, as introduced by Joseph et al. (2021). The A-OSE quantifies the number of unknown objects that are incorrectly classified as known (lower is better). In Table 9, we present the A-OSE values alongside the U-Recall metric for the Hierarchical Split.

hierarchical Split	Task 1		Task 2		Task 3	
	U-Recall (\uparrow)	A-OSE(\downarrow)	U-Recall (\uparrow)	A-OSE(\downarrow)	U-Recall (\uparrow)	A-OSE(\downarrow)
OW-DETR	7.0	42,540	11.0	26,527	8.8	20,034
PROB	29.4	14,962	43.9	8,929	52.7	5,387
Hyp-OW (Ours)	34.9	7,420	47.5	3,849	55.2	4,611

Table 9: **Unknown Confusion.** Hyp-OW achieves the highest U-Recall, indicating its superior ability to detect unknown objects. Additionally, it exhibits the lowest A-OSE, further emphasizing its capability to effectively differentiate between known and unknown objects.

C.2 Unveiling Co-occurrence Learning: The Influence of Bounding Box Overlap

In the Object Detection (OD) task, where bounding boxes can overlap, the model has a tendency to learn co-occurrence patterns of frequently appearing classes. This phenomenon is quantitatively demonstrated in the Heatmap Figure 6, where the model exhibits high similarity between classes that do not have any semantic relationship, such as ‘person’ and ‘tie’, or ‘teddy bear’ and ‘bicycle’ (highlighted in teal). This observation is further supported by Figures 7, 8 and 9, which provide qualitative examples of this co-occurrence learning.

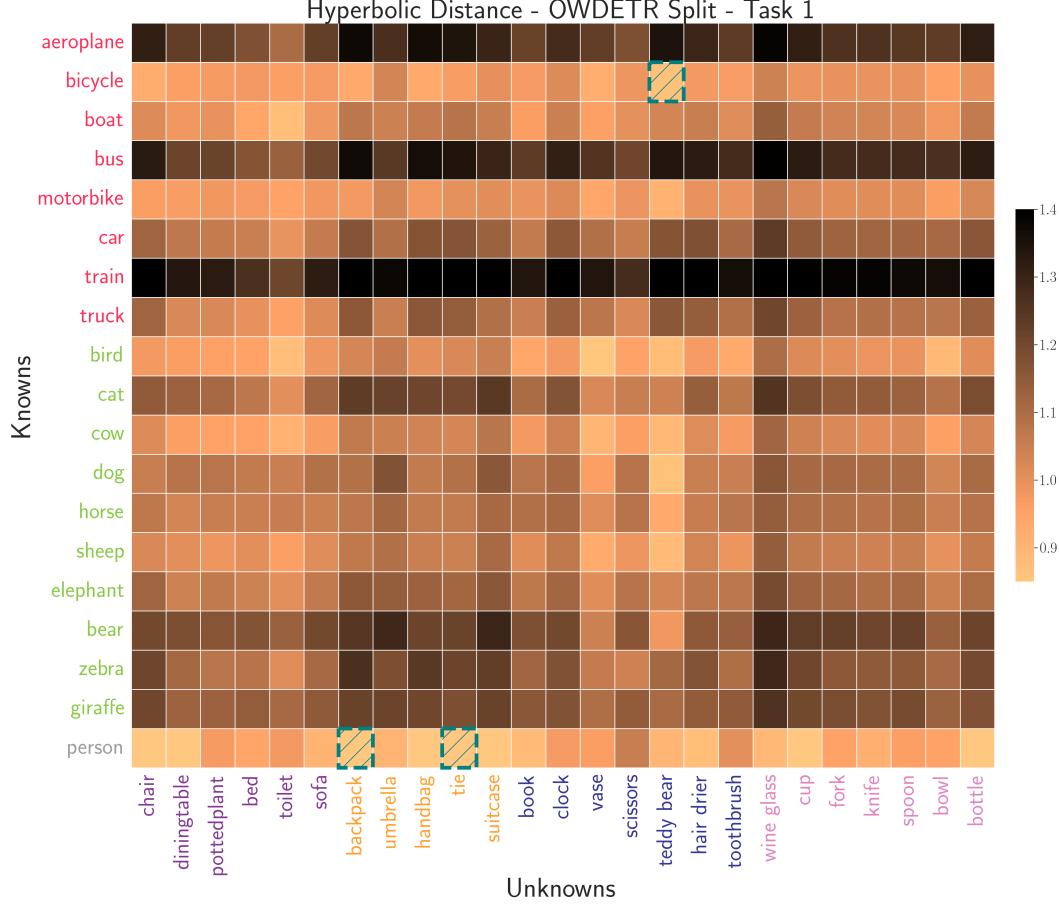


Figure 6: **Hyperbolic Knowns-Unknowns Distance Heatmap.** Lighter colors indicate lower hyperbolic distance and higher similarity. In the presence of images with high bounding box overlap, the model can learn frequent associations, such as ‘person’ and ‘tie’ or ‘backpack’ (teal hatched boxes, last row) or ‘teddy bear’ and ‘bicycle’.

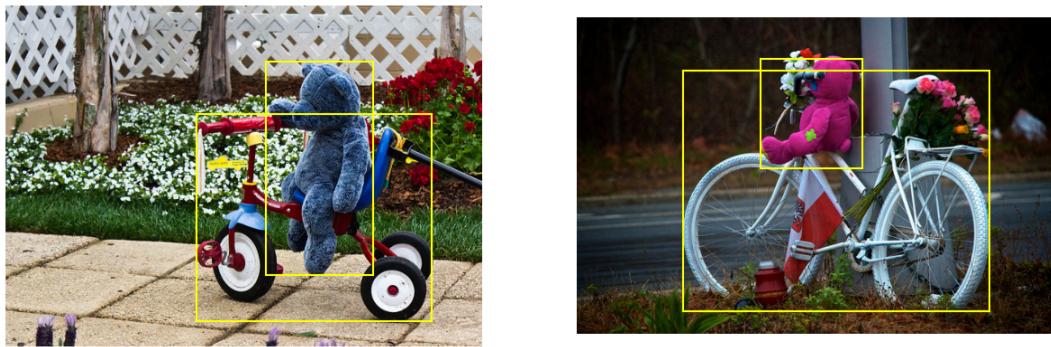


Figure 7: Bounding boxes overlap between classes 'teddy bear' and 'bicycle'.

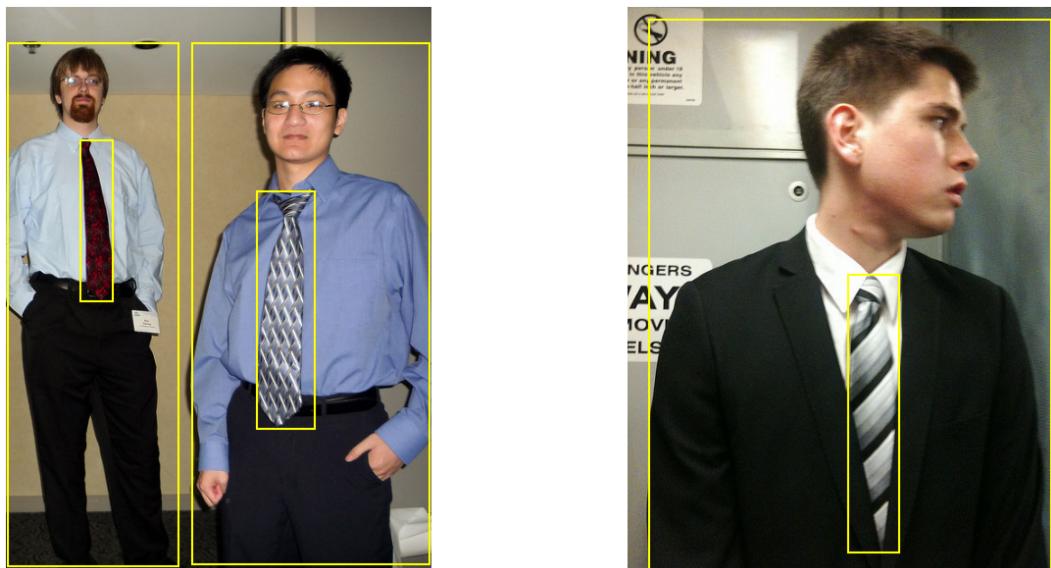


Figure 8: Bounding boxes overlap between classes 'person' and 'tie'.



Figure 9: Bounding boxes overlap between classes 'person' and 'backpack'.

D Ablation Analysis

In this section, we present comprehensive results of the ablation analysis. We then delve into a detailed discussion of the impact of each component.

D.1 Detailed Ablation Results

To analyze the impact of each component, we systematically deactivate individual modules and observe the resulting performance (Table 10). In summary, deactivating the Adaptive Relabeling module leads to a significant decrease in known class detection (mAP) while maintaining a stable U-Recall. On the other hand, the SuperClass Regularizer and Hyperbolic Distance modules are responsible for the U-Recall performance. These modules play a crucial role in achieving a balance between known class detection (mAP) and unknown class retrieval (U-Recall) maintaining an equilibrium in the detection process.

	Task 1			Task 2			Task 3			Task 4				
	U-Recall (\uparrow)	mAP (\uparrow)	Current	U-Recall (\uparrow)	mAP (\uparrow)	Both	U-Recall (\uparrow)	mAP (\uparrow)	Both	U-Recall (\uparrow)	mAP (\uparrow)	Both		
Hyp-OW (Ours)	34.9	49.9		47.5	42.0	49.0	45.5	55.2	44.4	44.1	44.3	42.8	47.0	43.9
w/Cosine Distance ($c = 0$)	32.8 (-2.1)	49.0 (-0.9)		46.4 (-1.1)	42.0	48.8	45.4	55.4	43.6	42.3	43.2 (-1.1)	42.4	45.2	43.1
w/o SuperClass Regularizer ($\beta = 0$)	32.0 (-2.9)	50.0		47.1	41.3	48.9	45.1	52.9 (-2.3)	43.8	43.5	43.7	42.8	45.9	43.5
w/o Adaptive Relabeling	34.7	41.2 (-8.7)		47.6	36.2	41.5	38.9 (-6.6)	54.1 (-1.1)	36.5	36.6	36.5 (-7.8)	34.5	37.9	36.1 (-7.8)

Table 10: Impact of each component of Hyp-OW on Hierachichal Split.

Deactivating Relabeling Module: To asses the impact of this module, we adopt PROB’s (Zohar et al., 2022) methodology where every unmatched query q^u is labelled as unknowns.

Deactivating SuperClass Regularizer: This is done by setting $\beta = 0$

Deactivating Hyperbolic Embedding: This is done by setting $c = 0$

D.2 Effectiveness of our Semantic Adaptive Relabeling Scheme

We now provide insights into the effectiveness of our Semantic Distance-based Relabeling Scheme. Figure 10 and 11 show the heatmap distance between known (left) and unknown items (bottom) for Task 1 in the OWOD and Hierarchical Split, respectively. Lighter colors indicate lower distances and higher similarity between known and unknown classes.

We can observe a significant similarity between known and unknown classes belonging to the same category, as highlighted by teal hatched boxes in both heatmaps. For example, in the OWOD Split, the animal classes from the knowns exhibit high similarity with the unknown classes. This is observed by comparing the classes on the left side of the heatmap, such as bird, cat, and cow, with the classes at the bottom, such as elephant and bear. In the Hierarchical Split, classes from the food category also demonstrate high similarity (see banana, apple versus cake and orange).

These findings emphasize the effectiveness of our Hyperbolic Distance-based Relabeling Scheme in capturing and leveraging the hierarchical and semantic relationships between known and unknown classes.

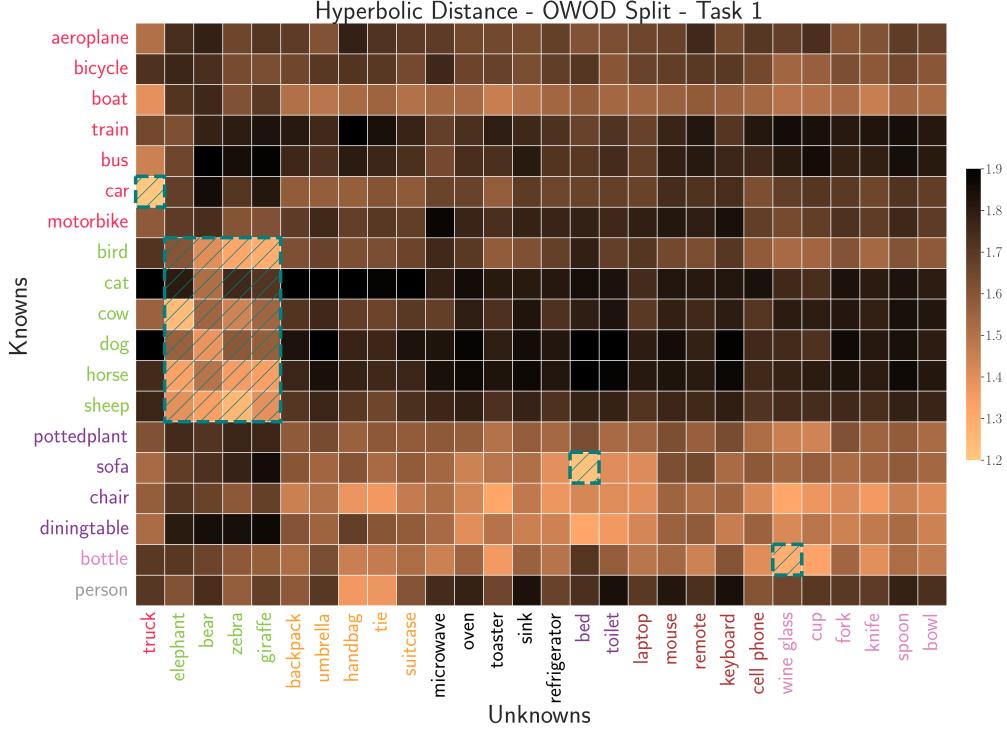


Figure 10: **Hyperbolic Knowns-Unknowns Distance Heatmap:** In the heatmap, lighter colors indicate higher similarity. We observe that classes belonging to the same category, such as 'car' and 'truck' or 'bed' and 'sofa', exhibit higher similarity.

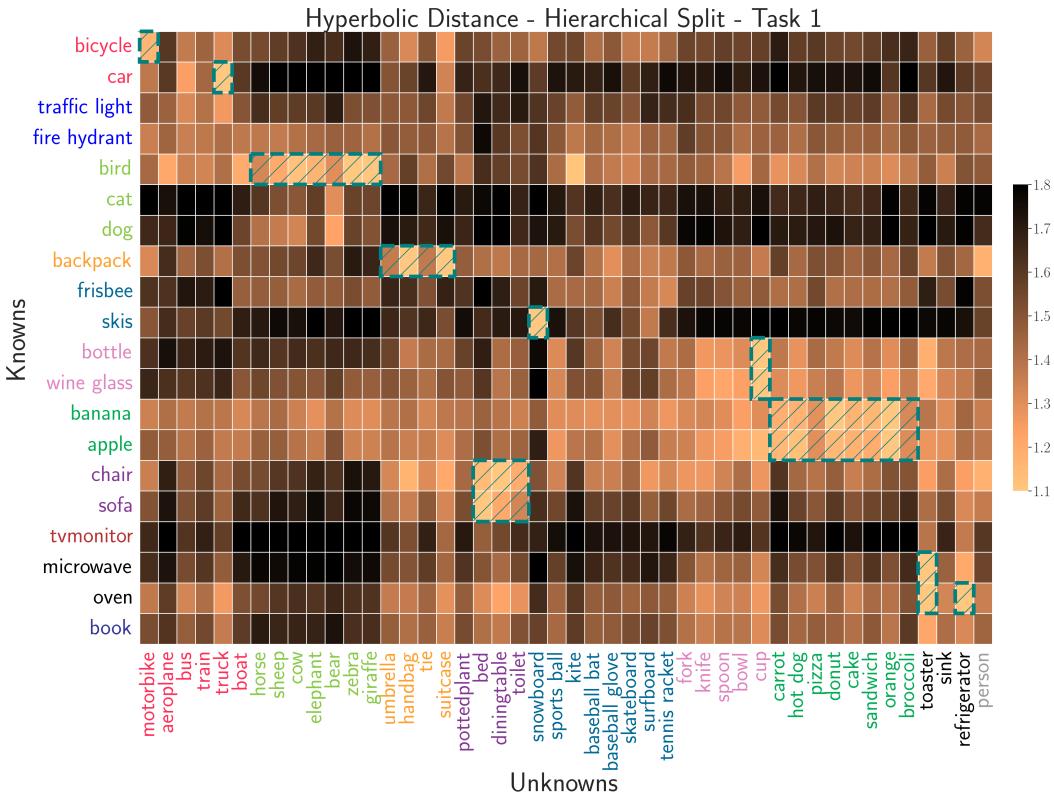


Figure 11: **Hyperbolic Knowns-Unknowns Distance Heatmap:** Lighter colors indicate higher similarity. We can observe that classes from the same category, such as 'skis' and 'snowboard' or 'toaster' and 'microwave', exhibit higher similarity.

D.3 SuperClass Regularizer

Figure 12 illustrates the Hyperbolic Distance between each SuperClass (left) and class (bottom) with (top) and without (bottom) our Regularizer. The inclusion of the Regularizer results in a wider range of values (from 0.7 to 2.30) compared to the absence of the Regularizer, where the range is narrower (from 0.78 to 1.2). This demonstrates that our Regularizer effectively pushes classes from different categories apart while bringing classes from the same category closer together. Additionally, without the Regularizer, the values are compressed, resulting in a squeezed inter-category distance, as evidenced by the lighter colors and weaker color contrast in the top plot. T-sne visualization (Figure 13) shows a qualitative visualization of this phenomenon.

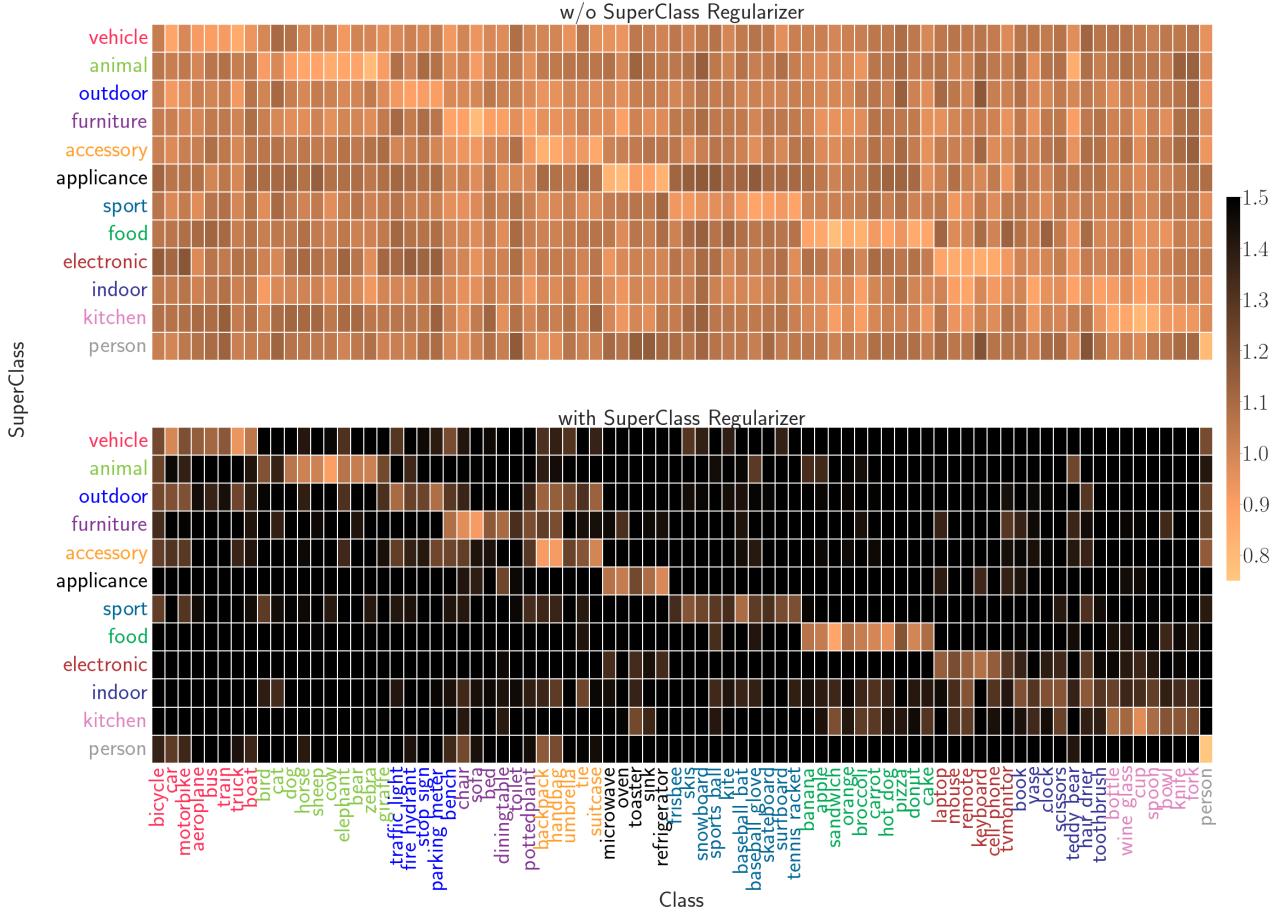


Figure 12: Hyperbolic SuperClass-Class Heatmap Distance.

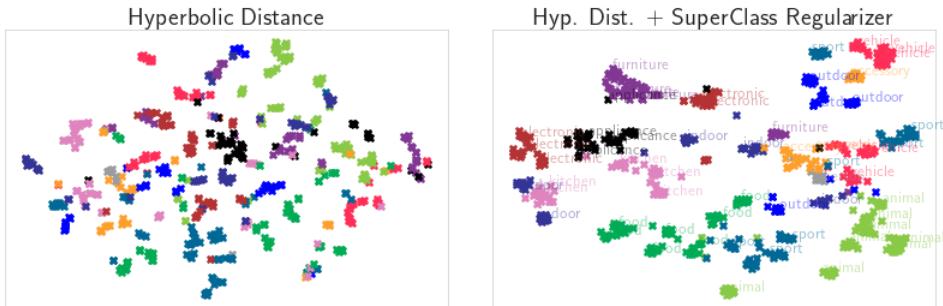


Figure 13: t-SNE plot of the learned class representations, with colors representing their respective categories.

D.4 Cosine versus Hyperbolic Distance

Figure 14 displays a t-SNE plot of the learned representation, where each class is color-coded based on its category. While both representations may initially appear scattered, the hyperbolic (right) distance-based representation exhibits improved grouping of classes from the same category. On the other hand, the cosine distance-based representation (left) lacks clear clustering properties between categories, as seen in the appliance category (black), animal category (light green), and furniture category (purple).

These observations highlight the advantage of utilizing hyperbolic distance for capturing meaningful category hierarchy relationship in the embedding space.

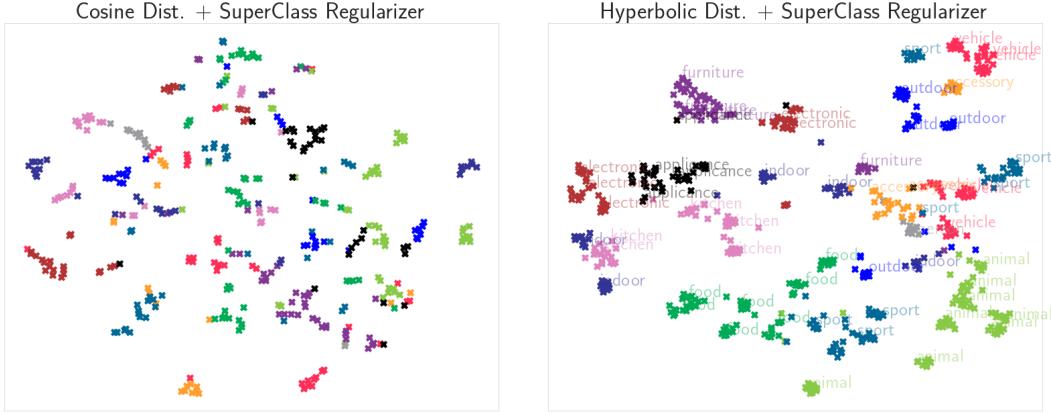


Figure 14: **t-SNE plot of the learned class embedding where each color represents the class’ category.**

D.5 Impact of Curvature Coefficient c

To investigate the influence of the curvature coefficient on the hyperbolicity of the embedding space, we conducted experiments using different values of c . Our focus was on Task 1 of the Hierarchical Split, where we assessed the model’s performance for four values of c : 0.0, 0.1, 0.2, and 0.5. The results are presented in Table 11. Notably, the curvature coefficient primarily impacts the U-Recall metric, while the mAP metric remains relatively stable. This observation is expected since the curvature coefficient c directly influences the learned hierarchical structure (hence the Semantic Similarity Distance), which is the cornerstone of our Adaptive Relabeling Scheme for unknown retrieval. Among the tested values, we found that $c = 0.1$ yielded the optimal performance, consistent with previous findings in the literature (Ermolov et al., 2022; Khrulkov et al., 2020b; Yue et al., 2023).

Hierarchical Split - Task 1

	U-Recall (\uparrow)	mAP(\uparrow)
$c = 0$ (Cosine Distance)	32.8	49.0
$c = 0.1$ (Hyp-OW)	34.9	49.9
$c = 0.2$	33.3	49.5
$c = 0.5$	32.3	49.8

Table 11: **Impact of curvature coefficient c .**