Hyperbolic Learning with Synthetic Captions for Open-World Detection

Fanjie Kong[1*],   Yanbei Chen[2†],   Jiarui Cai[2],   Davide Modolo[2]
[1]Duke University   [2]AWS AI Labs

fk43@duke.edu,   {yanbec,cjiarui,dmodolo}@amazon.com

Abstract

Open-world detection poses significant challenges, as it requires the detection of any object using either object class labels or free-form texts. Existing related works often use large-scale manual annotated caption datasets for training, which are extremely expensive to collect. Instead, we propose to transfer knowledge from vision-language models (VLMs) to enrich the open-vocabulary descriptions automatically. Specifically, we bootstrap dense synthetic captions using pre-trained VLMs to provide rich descriptions on different regions in images, and incorporate these captions to train a novel detector that generalizes to novel concepts. To mitigate the noise caused by hallucination in synthetic captions, we also propose a novel hyperbolic vision-language learning approach to impose a hierarchy between visual and caption embeddings. We call our detector "HyperLearner". We conduct extensive experiments on a wide variety of open-world detection benchmarks (COCO, LVIS, Object Detection in the Wild, RefCOCO) and our results show that our model consistently outperforms existing state-of-the-art methods, such as GLIP, GLIPv2 and Grounding DINO, when using the same backbone.

1. Introduction

An intelligent perception model of our visual world requires the open-world generalizability of understanding any object. To enable such generalizability, recent advances in object detection integrate vision models with natural language models to enrich their open-world knowledge [15, 24, 50, 60, 63, 65]. Most of these works [15, 50, 60, 65] consider the open-vocabulary setting in which a novel object is specified by its class label (e.g., *panda*). Other works [24, 63] cast detection as a generalized localization task and consider a more challenging open-set setting in which objects are described by free-form texts (e.g., *a black-and-white giant bear*). In this work, we tackle the task of detecting and localizing any object using both class labels and
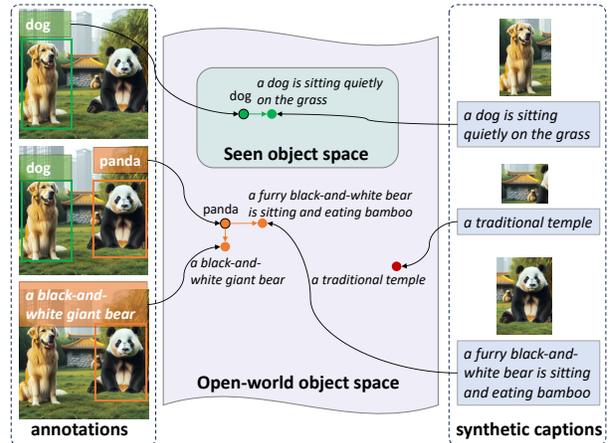
Figure 1. We tackle the task of detecting seen and unseen objects using keywords (e.g., *panda*) or free-form texts (e.g., *a black-and-white giant bear*) in open world. We exploit synthetic captions from pre-trained caption models to bring rich open-world knowledge for training. As synthetic captions may be noisy, we propose to align visual features with text embeddings in a structural hierarchy to learn robustly and effectively from these captions.

free-form texts (Figure 1). We refer to this challenging task as "open-world detection".

To improve generalization in open world, existing detection methods [29, 60, 63] utilize language descriptions from large-scale image captioning datasets (e.g., MS-COCO captions [32], conceptual captions [46]) for training. These descriptions are however expensive to manually annotate [32] or design human-crafted data cleaning pipeline to process alt-text [46]. To overcome this limitation, we propose to leverage recent advanced vision-language models to bootstrap machine-generated captions automatically. We collect these synthetic captions on different regions of an image to generate a dense understanding of all its concepts: from rich semantic information of known/seen objects (e.g., their attributes, states, actions and interactions) to diverse open-vocabulary descriptions of new concepts. Our intuition is that by aligning the detector features with these rich semantics, we can boost the model capability of understanding and recognizing any object, including novel ones.

---

*Work done during an internship with AWS AI Labs.
†Corresponding author.

Nevertheless, one needs to be careful when using synthetic captions, as they are sometimes hallucinated or unrelated to the objects in the image [44]. For instance, given an image of a *panda* (Figure 1), the model may generate the caption: "*a furry black-and-white bear is sitting and eating bamboo*", hallucinating the action "*eating bamboo*", which usually co-occurs with the animal *panda*. To mitigate the negative impact of aligning visual features with noisy captions, we propose to learn visual-textual alignment in a hierarchical structure, in which the synthetic caption entails the keyword of object, along with other semantics. We refer to this hierarchy structure as *'caption entails object'*. To impose such a structure in our representation space, we propose a novel hyperbolic vision-language learning objective which ensures that the caption embeddings entails the visual object embeddings in the representation space. Our formulation is inspired by the advances of structural representation learning with hyperbolic geometry [8, 9, 12, 13, 22]. Rather than improving the unimodal representation learning on purely images [13] or texts [9], our hyperbolic learning objective is introduced for cross-modal learning on images and synthetic captions, and especially addresses the hallucination problem of synthetic captions to advance the open-world generalization in detection.

To summarize, our contribution is three folds:

(1) We propose to boost open-world detection using synthetic captions generated by powerful pre-trained VLMs. These synthetic captions provide rich language descriptions of both seen and unseen objects, enabling the detection of both new class labels and free-form texts.

(2) We formulate a novel hyperbolic vision-language learning objective to help learning robust representation using noisy synthetic captions. Our hyperbolic loss formulation imposes a meaningful structural hierarchy in the representation space to specifically tackle the hallucination problem in these captions.

(3) We show that our approach achieves the state-of-the-art performance on a variety of detection and localization datasets in the open-world setting, including COCO, LVIS, Object Detection in the Wild (ODiW), and Ref-COCO. Our approach outperforms existing open-world detectors such as GLIP, GLIPv2, and Grounding DINO when using the same backbone and datasets for training.

2. Related Work

Open-vocabulary detection (OVD). Standard detection models like Faster R-CNN [14] and DETR [1] succeed in detecting objects from a fixed vocabulary, but are not capable of localizing novel concepts. To overcome this limitation, OVD methods take advantage of extensive image-text datasets through vision-language models (VLMs). OVD approaches follow two trends: learning novel object concepts from image-level supervision [65], and transferring

knowledge from VLMs, which includes distilling visual-semantic cues [15, 50], aligning image-text representations [51, 59, 64], and exploiting pseudo-labels from captioning models [4, 28]. In line with the more contemporary trend, our approach seeks to assimilate novel semantic concepts by probing VLMs. Specifically, we focus on distilling knowledge from large amount of (noisy) captions generated by these VLMs. Moreover, we also delve into the integration of free-form texts in open-world localization, lifting the model's proficiency in understanding complex referring expressions beyond plain object classes.

Vision-language models (VLMs) in detection. VLMs have demonstrated remarkable abilities in tasks like image classification and image-text retrieval [7, 19, 41, 52, 55, 61]. Recently, VLMs have also pioneered the state-of-the-art in detection, as represented by GLIP [29, 63], Region-CLIP [64] and Grounding DINO [34]. These methods use large-scale image-text datasets to train the detector, thus enhancing the knowledge of both seen and unseen concepts. Concurrently, other studies [3, 15, 50, 65] have focused on exploiting the visual-semantic knowledge from VLMs, e.g., Detic [65] and ScaleDet [3] use the text encoder from CLIP to encode the class labels of any seen and unseen objects. Our work contributes to this evolving direction with a novel hyperbolic vision-language learning approach to learn with synthetic captions from VLMs, which avoids the need of expensive manual annotated image-text datasets.

Learning with captions. Image captions convey rich information about objects in an image, such as object attributes, actions, relationships and scenes. Recent works have highlighted the advantages of harnessing caption data to enhance visual representation learning [7, 11, 48, 53, 56]. The majority of these studies exploit *manual annotated captions*, training visual encoders alongside auxiliary caption generators to learn more generalized visual features [7, 48, 56]. A few studies investigate the use of *machine generated captions* to augment supervision for visual learning [11, 53]. Our approach takes an innovative step further by employing synthetic captions to improve open-world generalization in detection. Importantly, we tackle the hallucination issue that often presents in synthetic captions by introducing a novel hyperbolic loss formulation.

Hyperbolic learning. Representation learning based on hyperbolic geometry has shown considerable success in various domains such as text embeddings [9], text generation [5] and knowledge graph modeling [2], as the hyperbolic geometry could model the intrinsic hierarchy among data. Recently, a few works explore hyperbolic learning in vision tasks to model the hierarchical relation between objects and scenes [13], or images and texts [8]. Our work is built on the foundation of hyperbolic learning to advance performance in detection. We specially delve into modelling the hierarchy between visual objects and synthetic

captions, and formulate a novel hyperbolic learning objective to learn upon such hierarchy.

3. Proposed Approach

We first present the preliminaries of our model in §3.1. We then introduce our caption bootstrapping strategy which provides rich open-vocabulary descriptions for training (§3.2). Finally, we present our hyperbolic vision-language learning approach which exploits rich semantics from synthetic captions to boost open-world generalization (§3.3).

3.1. Preliminaries on model architecture

Training an open-world detector. Object detectors are trained to predict a list of bounding box location $B_i \in \mathbf{R}^4$ and class label $L_i \in \mathbb{R}^n$. They often consist of an visual encoder, a bounding box regressor and a visual classifier. During training, a bounding box regression loss $\mathcal{L}_{bbox}$ and a classification loss $\mathcal{L}_{cls}$ are jointly optimized to predict the object bounding box location and class label, i.e.,

$$\mathcal{L}_{Det} = \mathcal{L}_{bbox} + \mathcal{L}_{cls}. \tag{1}$$

Most detectors can be categorized into one-stage [47, 67] or two-stage [18, 42, 66]. Detectors in the latter set employ an additional loss to train a region proposal network ($\mathcal{L}_{rpn}$) for predicting the objectness score of each region proposal. For the classification loss $\mathcal{L}_{cls}$, classic detectors often adopt a cross-entropy loss [42], which works well for classifying known classes, but it cannot model novel classes. To overcome this limitation, open-world detectors [29, 65] leverage text embeddings to represent the class labels. In details, the classification loss $\mathcal{L}_{cls}$ helps aligning the visual feature embedding v_i with its class label embedding l_i, i.e.,

$$\mathcal{L}_{cls} = -\log \frac{\exp(\mathrm{sim}(v_i, l_i)/\tau)}{\sum_{j=1}^n \exp(\mathrm{sim}(v_i, l_j)/\tau)}, \tag{2}$$

where v_i is visual feature embedding from the detector, $\mathrm{sim}(\cdot, \cdot)$ measures cosine similarity between embeddings, τ is the temperature. The class label embedding l_j is extracted with a pre-trained text encoder from CLIP [41]. At test time, the predicted label is $L_i = \arg\max_j \{\mathrm{sim}(v_i, l_j)\}_{j=1}^n$, where n is the number of pre-defined class labels (seen during training). One can augment these n classes with novel classes (by adding their class label embeddings) to enable the open-set recognition on these novel concepts.

Cross-modal attention. The above formulation (Eq. (2)) enables the detection of novel objects using their class label embeddings. However, it does not capture the spatial awareness of language semantics on different visual objects, thus fails to localize objects with more precise information, e.g., localizing *"the person playing a bass guitar"* in Figure 2. To achieve open-world detection with free-form texts, we propose to inject the language semantics into

the visual feature embedding v_i by a cross-modal attention module [27, 29]. Specifically, given any text embeddings $T = \{t_j\}_1^n$, we compute the cross-modal attention between v_i, T as:

$$
\begin{aligned}
v_i^{(q)} &= (W^{(q)})^\top v_i, \ T^{(k)} = TW^{(k)}, \ T^{(v)} = TW^{(v)}, \\
A_i &= T^{(k)} v_i^{(q)}/\sqrt{d}, \ v_i^l = \mathrm{softmax}(A_i) \, T^{(v)} W^{(out)},
\end{aligned} \tag{3}
$$

where $W^{(q)}$, $W^{(k)}$, $W^{(v)}$ are the trainable query, key, and value matrices in multi-head attention [49], d is the dimension of hidden embeddings $v_i^{(q)}$ and $T^{(k)}$, and v_i^l is the visual feature embedding fused with language semantics. To further inject spatial awareness into the visual feature embedding, we add a positional encoding layer [10] to fuse the visual feature embedding v_i with region proposal feature p_i:

$$v_i^s = \mathrm{PositionalEncoding}(v_i, p_i). \tag{4}$$

Finally, to attain a spatial-aware visual feature embedding fused with language semantics, we combine the language-aware and spatial-aware visual embeddings v_i^l (Eq. (3)), v_i^s (Eq. (4)) with the region proposal network $\mathrm{RPN}(\cdot)$:

$$\tilde{v}_i = \mathrm{RPN}(v_i^l + v_i^s). \tag{5}$$

Within this formulation, our detector can localize objects given any texts by replacing the original visual embedding v_i with the new visual embedding $\tilde{v}_i$, and compute the localization score based on similarity between $\tilde{v}_i$ and the text embeddings $\{t_j\}_1^n$, i.e., $Loc_i = \arg\max_j \{\mathrm{sim}(\tilde{v}_i, t_j)\}_{j=1}^n$.

3.2. Bootstrapping synthetic captions

To enable generalization on novel objects, existing methods often exploit manual annotated datasets such as image classification [43], captioning [46] and grounding dataset [23] to provide knowledge of novel concepts. Rather than relying on expensive manual annotations, we propose to bootstrap synthetic captions from a pre-trained VLM. Similar to manual annotations, these captions also provide rich open-vocabulary descriptions of novel concepts. In this paper, we use BLIP2 [28] to generate captions – which is trained on a large-scale dataset of 129M images. It is worth noting that our model is agnostic to the VLM used for caption generation, and thus can benefit from more powerful VLMs.

Region sampling for caption bootstrapping. We collect three sets of region proposals. First, for seen/known classes we use the ground truth bounding boxes $\mathcal{B}$. Then, to attain regions on novel objects, we propose to use two region sampling strategies: *proposal sampling* and *grid sampling*. In *proposal sampling*, we sample region proposals with high objectness scores and apply non-maximum suppression to remove duplicates; this leads to the set of regions $\mathcal{P}$. In *grid sampling*, we split the image into $k \times k$ grids similar to ViT [10] to get a set of region crops $\mathcal{G}$. Together, these produce a
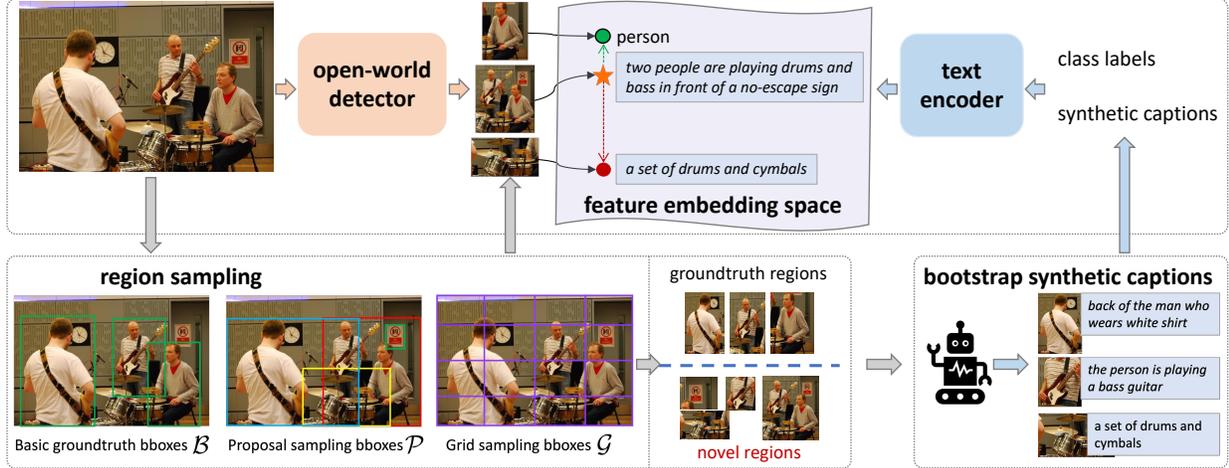
Figure 2. Approach overview. Given an image, our open-world detector extracts visual feature embeddings on region crops, and aligns visual feature embeddings with text embeddings extracted from a pre-trained text encoder (§3.1). To obtain synthetic captions for both seen and novel objects, we propose region sampling to augment region crops, and adopt a pre-trained image captioner to bootstrap synthetic captions on these region crops (§3.2). To learn from synthetic captions effectively, we propose to align visual feature embeddings and caption embeddings in a structural hierarchy through hyperbolic vision-language learning in the hyperbolic space (see Figure 3, §3.3).

rich set of region crops: $\mathcal{B} \cup \mathcal{P} \cup \mathcal{G}$ that encapsulate both seen and unseen objects. Finally, we generate synthetic captions $\mathcal{C}$ densely over these region crops with the BLIP2 captioner:

$$\mathcal{C} = \text{Captioner}_{\text{BLIP2}}(\mathcal{B} \cup \mathcal{P} \cup \mathcal{G}), \quad (6)$$

where the captions $\mathcal{C}$ are then used for model training.

3.3. Hyperbolic vision-language learning

Vanilla contrastive learning. Our bootstrapping synthetic captions provide diverse language descriptions on both seen and unseen objects in different image regions, thus offering rich knowledge to boost open-world generalization. To learn from such knowledge, we can adopt the vanilla *contrastive loss* [41] for aligning the visual feature embedding $\tilde{o}_i$ (Eq. (5)) with the captions embedding c_i attained in (6):

$$\mathcal{L}_{cap} = -\log \frac{\exp(\text{sim}(\tilde{v}_i, c_i,)/\tau)}{\sum_{j=1}^{m} \exp(\text{sim}(\tilde{v}_i, c_j)/\tau)}, \quad (7)$$

where $\mathcal{L}_{cap}$ is computed on a batch of m caption embeddings, which enables learning about the novel concepts described in the synthetic captions. However, synthetic captions suffer from the hallucination problem [44], where concepts unrelated to the image are added to the generated caption; this simple contrastive loss is not equipped to handle them and its learning suffers from the generated noise. To learn more effectively from these captions, we propose to align visual and text embeddings in a hierarchy.

Hyperbolic contrastive learning. The synthetic caption on a region often entails the object descriptions along with other content. Thus, we argue that the object and caption should follow a *'caption entails object'* hierarchy in a tree-like structure (Figure 3). Inspired by recent advances in hyperbolic learning to model the partial order of *'scene entails*

(a) Euclidean space   **(b)** Hyperbolic space
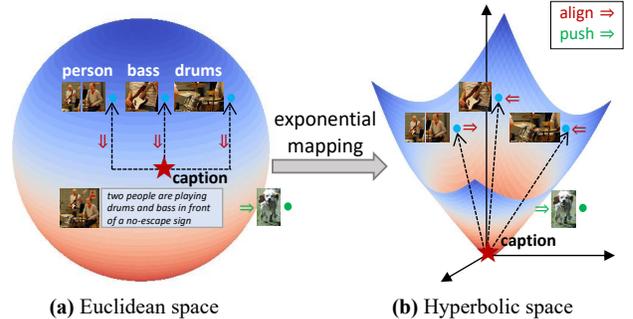
Figure 3. Illustration of hyperbolic vision-language learning. The visual and caption embeddings are lifted from **(a)** Euclidean space to **(b)** Hyperbolic space by exponential mapping (Eq. (8)). To learn the partial order of *'caption entails object'*, we propose hyperbolic contrastive loss, hyperbolic entailment loss (Eq. (10), Eq. (13)) to align visual and caption embeddings in hierarchy, thus ensuring the hallucination in caption is not aligned directly with visual embeddings to negatively impact the model learning.

object' [13] or *'text entails image'* [8], we propose to model our *'caption entails object'* hierarchy in a hyperbolic space.

Specifically, we introduce our hyperbolic contrastive learning based on the Lorentzian distance in hyperbolic geometry [25], which lifts the embeddings from Euclidean space to the Lorentz hyperboloid to represent the structural hierarchy. Formally, we first project the visual feature embedding $\tilde{v}_i$ and caption embedding c_i to the hyperboloid with exponential mapping (Eq. (8)), and then compute their Lorentzian distance on the hyperboloid (Eq. (9)) as follows:

$$v_i^{\mathcal{H}} = \text{expm}_{\mathbf{0}}(\tilde{v}_i), \ c_i^{\mathcal{H}} = \text{expm}_{\mathbf{0}}(c_i),$$

$$\text{with } \text{expm}_{\mathbf{0}}(x) = \frac{\sinh(\sqrt{C}\|x\|)}{\sqrt{C}\|x\|}, \quad (8)$$

where $\exp m_{\mathbf{0}}(\cdot)$ is an exponential mapping, and C is a learnable curvature parameter. Given two mapped embeddings $v_i^{\mathcal{H}}, c_i^{\mathcal{H}}$, we follow existing Lorentz model [8, 25, 26, 40], and compute the Lorentzian distance on the hyperboloid as:

$$d_{\mathcal{H}}(\tilde{v}_i, c_i) = \sqrt{1/C} \cdot \cosh^{-1}(-c\langle \tilde{v}_i, c_i \rangle_{\mathcal{H}}),$$
$$\text{with } \langle \tilde{v}_i, c_i \rangle_{\mathcal{H}} = \langle \tilde{v}_i, c_i \rangle - \sqrt{1/C + \|\tilde{v}_i\|}\sqrt{1/C + \|c_i\|}, \quad (9)$$

where $\langle \cdot, \cdot \rangle_{\mathcal{H}}$ is the Lorentz inner product induced by the Riemannian metric of the Lorentz model [8, 25, 26, 40], and $d_{\mathcal{H}}(\cdot, \cdot)$ is the distance between two embeddings on the hyperboloid. By mappings the embeddings to the hyperbolic space, we derive the *hyperbolic contrastive loss* by replacing cosine similarity measure $\text{sim}(\cdot, \cdot)$ in Eq. (7) with the hyperbolic distance measure $d_{\mathcal{H}}(\cdot, \cdot)$, i.e,

$$\mathcal{L}_{cap}^{\mathcal{H}} = -\log \frac{\exp(-d_{\mathcal{H}}(v_i^{\mathcal{H}}, c_i^{\mathcal{H}})/\tau)}{\sum_{j=1}^{B} \exp(-d_{\mathcal{H}}(v_i^{\mathcal{H}}, c_j^{\mathcal{H}})/\tau)}, \quad (10)$$

where $\mathcal{L}_{cap}^{\mathcal{H}}$ is our loss function to align visual embedding and caption embedding in the hyperbolic space.

Hyperbolic entailment. To impose our proposed structural hierarchy of *'caption entails object'*, we formulate a hyperbolic entailment loss based on the property of an *entailment cone* – which defines a cone in the hyperbolic space to enforce partial order between embeddings [26]. Specifically, we define an entailment cone for caption embedding $c_i^{\mathcal{H}}$ as:

$$A(c_i^{\mathcal{H}}) = \sin^{-1}\left(\frac{2K}{\sqrt{C}\|c_i^{\mathcal{H}}\|}\right) \quad (11)$$

where K is a constant 0.1 used to avoid numerical overflow; $A(\cdot)$ is half aperture of the cone in hyperbolic space. With $A(c_i^{\mathcal{H}})$, we introduce the entailment constraints to impose the hierarchy of *'caption entails object'*, by ensuring (1) the visual object embedding $v_i^{\mathcal{H}}$ lies inside the cone, and (2) other visual embeddings $v_j^{\mathcal{H}}$ ($j \neq i$) lie outside the cone.

Formally, we denote the exterior angle between $v_i^{\mathcal{H}}$ and $c_i^{\mathcal{H}}$ as $\angle(c_i^{\mathcal{H}}, v_i^{\mathcal{H}})$[1], and introduce a loss term to ensure the entailment constraint (1):

$$E(c_i^{\mathcal{H}}, v_i^{\mathcal{H}}) = \max(0, \angle(c_i^{\mathcal{H}}, v_i^{\mathcal{H}}) - A(c_i^{\mathcal{H}})), \quad (12)$$

where $E(\cdot, \cdot)$ ensures $\angle(c_i^{\mathcal{H}}, v_i^{\mathcal{H}}) \leqslant A(c_i^{\mathcal{H}})$, thus enforcing $v_i^{\mathcal{H}}$ to lie inside the cone of $c_i^{\mathcal{H}}$. To further ensure both entailment constraints (1) and (2), we introduce the following *hyperbolic entailment loss* with two loss terms:

$$\mathcal{L}_{\text{entail}} = E(c_i^{\mathcal{H}}, v_i^{\mathcal{H}}) + \sum_{j \neq i} \max(0, \gamma - E(c_i^{\mathcal{H}}, v_j^{\mathcal{H}})), \quad (13)$$

---

[1]Note: the exterior angle $\angle(a, b)$ in hyperbolic space is computed as $\angle(a, b) = \cos^{-1}\left(\frac{\sqrt{1/c + \|b\|} + \sqrt{1/c + \|a\|}c\langle a_i^{\mathcal{H}}, b_i^{\mathcal{H}}\rangle_{\mathcal{H}}}{\|a_i^{\mathcal{H}}\|\sqrt{c(\langle a_i^{\mathcal{H}}, b_i^{\mathcal{H}}\rangle)^2 - 1}}\right)$, c is curvature.

where the first term is Eq. (12); the second term is a max margin loss to ensure $\angle(c_i^{\mathcal{H}}, v_j^{\mathcal{H}}) \geqslant A(c_i^{\mathcal{H}}) + \gamma$, thus enforcing $v_j^{\mathcal{H}}$ to lie outside the cone of $c_i^{\mathcal{H}}$ with a margin γ.

Overall objective. Finally, given our proposed hyperbolic loss functions (Eq. (7), Eq. (13)), we formulate our learning objective as follows:

$$\mathcal{L}_{hyper} = \mathcal{L}_{bbox} + \mathcal{L}_{cls} + \mathcal{L}_{cap}^{\mathcal{H}} + \mathcal{L}_{\text{entail}}, \quad (14)$$

where $\mathcal{L}_{cap}^{\mathcal{H}}, \mathcal{L}_{\text{entail}}$ are our hyperbolic contrastive loss, and hyperbolic entailment loss to align visual and caption embedding in a structural hierarchy. An alternative to learn from synthetic captions is the vanilla contrastive loss (Eq. (7)), which leads to the baseline objective below:

$$\mathcal{L}_{baseline} = \mathcal{L}_{bbox} + \mathcal{L}_{cls} + \mathcal{L}_{cap}. \quad (15)$$

We compare our **Hyper**bolic **Learn**ing objective $\mathcal{L}_{hyper}$ (Eq. (14)) to alternative objectives $\mathcal{L}_{Det}$ (Eq. (1)), $\mathcal{L}_{baseline}$ (Eq. (15)) in ablation study to verify our design rationales. We refer our approach as "**HyperLearner**" in experiments.

4. Experiments

In the following, we detail our experimental settings (§4.1) and provide thorough evaluation of our approach on a wide set of benchmark datasets in comparison to the state-of-the-arts (§4.2). Finally, we provide insightful ablation study (§4.3) and qualitative results (§4.4) to analyze our approach.

4.1. Experimental settings

Implementation details. Our approach utilizes Center-NetV2 [65] as the first-stage region proposal network, succeeded by a cross-modal attention module as detailed in §3.1. Our text encoder is the 12-layer text transformer from CLIP ViT/B16 [41]. We used BLIP2 6.7b [28] to generate synthetic captions on region crops offline (§3.2). For model pre-training, we used two datasets: Object 365 (O365) [45] and GoldG [20], to ensure a fair comparison with other methods on the same evaluation benchmark datasets. In all experiments, we use Swin-Tiny [35] as the backbone, and undergo training for 100,000 iterations with a batch size of 64 and a learning rate 0.0001. The initial 10,000 iterations serve as a warm-up phase, which is essential for optimal convergence. We train our model on 8 A100 GPUs. More details are provided in the supplementary material.

Object detection benchmark datasets. We evaluate the open-world detection performance of our method on the most popular object detection datasets: (1) COCO [32], which has 80 common object classes, and (2) LVIS [17], which has 1203 object classes. Furthermore, we also evaluate on Object Detection in the Wild (ODinW) [29] – an assembly of 13 datasets, each one representing a different real-world fine-grained domain. Following the common

	Method	Backbone	#Params	FLOPs	Pre-training Data	COCO2017 val Zero-shot	Fine-tuning
1	Faster-RCNN [14]	RN50-FPN	42M	180G	COCO	-	40.2
2	Faster-RCNN [14]	RN101-FPN	54M	313G	COCO	-	42.0
3	Deformable DETR(DC5) [68]	RN50	41M	187G	COCO	-	41.1
4	CenterNetv2 [66]	RN50	76M	288G	COCO	-	42.9
5	Dyhead-T [6]	Swin-T	232M	361G	O365	43.6	53.3
6	GLIP-T(A) [29]	Swin-T	232M	488G	O365	42.9	52.9
7	GLIP-T(B) [29]	Swin-T	232M	488G	O365	44.9	53.8
8	GLIP-T(C) [29]	Swin-T	232M	488G	O365, GoldG	46.7	55.1
9	DINO-T [62]	Swin-T	-	-	O365	46.2	56.9
10	Grounding-DINO-T[1] [34]	Swin-T	172M	464G	O365	46.7	56.9
11	Grounding-DINO-T[2] [34]	Swin-T	172M	464G	O365, GoldG	48.1	57.1
12	Grounding-DINO-T[3] [34]	Swin-T	172M	464G	O365, GoldG, Cap4M	48.4	57.2
13	HyperLearner (Ours)	Swin-T	**90M**	**324G**	O365	**47.6**	**56.8**
14	HyperLearner (Ours)	Swin-T	**90M**	**324G**	O365, GoldG	**48.4**	**57.4**

Table 1. Comparison on COCO benchmark. Results are given on both zero-shot and fine-tuning settings. Metric: mAP.

Method	Pre-training Data	LVIS minival AP	APr	APc	APf
MDETR [20]	GoldG, RefCOCO	24.2	20.9	24.9	24.3
DETCLIP-T [54]	O365	28.8	26.0	28.0	30.0
GLIP-T (C) [29]	O365, GoldG	24.9	17.7	19.5	31.0
GLIP-T [29]	O365, GoldG, Cap4M	26.0	20.8	21.4	31.0
Grounding-DINO-T [34]	O365, GoldG	25.6	14.4	19.6	32.2
Grounding-DINO-T [34]	O365, GoldG, Cap4M	27.4	20.8	21.4	31.0
HyperLearner (Ours)	O365	25.5	25.9	27.5	23.7
HyperLearner (Ours)	O365, GoldG	**31.3**	**30.7**	**32.6**	**30.3**

Table 2. Comparison on LVIS benchmark. Metric: mAP.

practices, for zero-shot setting we evaluate our pre-trained models directly, while for fine-tuning setting, we fine-tune our pre-trained models on the target dataset. We report mean Average Precision (mAP) at IoU threshold of 0.5.

Object localization benchmark datasets. We evaluate our method on referring expression comprehension which aims to locate novel objects using natural language expressions. We employ the following benchmark datasets: Ref-COCO [21], RefCOCO+ [58] and RefCOCOg [37] (collectively referred as RefCOCO/+/g) to evaluate localizing the objects given descriptive free-form texts. All three datasets utilize the images from the original COCO dataset; however, they differ in their textual descriptions: RefCOCO+ emphasizes purely on appearance, excluding spatial reference found in RefCOCO; RefCOCOg provides more comprehensive descriptions in sentences rather than phases. Our evaluation adheres to the established evaluation protocol, employing top-1 accuracy to determine the successful localization of objects based on the description, in line with the metric used in previous works [20, 34, 57].

4.2. Comparison to SOTA methods

Comparison on COCO. In Table 1, we compare our HyperLearner to several recent models: (1) traditional two-stage object detectors, including Faster-RCNN [42], Deformable DETR [68] and CenterNetV2 [65]; (2) closed-set detectors which are adaptable for evaluation on COCO through class mapping between O365 and COCO, including Dyhead [6] and DINO [62]; (3) open-world detectors including GLIP [29] and Grounding DINO [34], which not only use image-text data for pre-training, but also utilize model architectures to fuse visual-language representations.

Table 1 details the comparison on model efficiency and performance on COCO. As shown, our model has much smaller parameter size and computation cost (measured in FLOPs), which significantly improve model efficiency for real-world deployment. Moreover, while being lightweight, our model achieves better performance in both zero-shot and fine-tuning settings, with an mAP of 48.4 and 57.4 as compared to 48.1 and 57.1 obtained by the best competitor Grounding DINO-T when using O365, GoldG for training. Notably, our model pre-trained on O365, GoldG (in row 14) obtains similar performance as Grounding DINO-T pre-trained on O365, GoldG and Cap4M, an additional large-scale manually-annotated caption dataset [29] (row 12). This clearly shows how powerful machine-generated captions can be when used properly in HyperLearner.

Comparison on LVIS. Table 2 shows the results on LVIS minival validation set. Our model demonstrates a significant advantage over GLIP and Grounding DINO on detecting novel rare classes – outperforming them by +13.0% and +16.3% in APr respectively, when pre-trained on O365 and GoldG. Our improved gain is maintained at 9.9% even when other methods benefit from the additional Cap4M dataset. The results suggest our proposed method's capability to detect novel and rare concepts, attributing to the effective incorporation of semantic knowledge via bootstrapping synthetic captions and the vision-language alignments brought by our novel hyperbolic loss formulation.

Comparison on ODinW. The ODinW dataset exemplifies more fine-grained, challenging real-world application scenarios. In Table 4, we report the average AP over its 13 subsets. Our finding aligns with those from the COCO and LVIS datasets, indicating that our model performs well in

	Method	Pre-training Data	Fine-tuning	RefCOCO			RefCOCO+			RefCOCOg	
				val	testA	testB	val	testA	testB	val	test
1	CNN-LSTM [37]	RefC	-	-	63.15	64.21	-	48.73	42.13	62.14	-
2	MAttNet [57]	RefC	✓	76.65	81.14	69.99	65.33	71.62	56.02	66.58	67.27
3	RefTR [30]	VG	✓	85.65	88.73	81.16	77.55	82.26	68.99	79.25	80.01
4	MDETR [20]	GoldG,RefC	✓	86.75	89.58	81.41	79.52	84.09	70.62	81.64	80.89
5	DQ-DETR [33]	GoldG,RefC	✓	88.63	91.04	83.51	81.66	86.15	73.21	82.76	83.44
6	GLIP-T(B) [29]	O365,GoldG	×	49.96	54.69	43.06	49.01	53.44	43.42	65.58	66.08
7	GLIP-T(C) [29]	O365,GoldG,Cap4M	×	50.42	54.30	43.83	49.50	52.78	44.59	66.09	66.89
8	Grounding-DINO-T [34]	O365,GoldG	×	50.41	57.24	43.21	51.40	57.59	45.81	67.46	67.13
9	Grounding-DINO-T [34]	O365,GoldG,RefC	×	73.98	74.88	59.29	66.81	69.91	56.09	71.06	72.07
10	Grounding-DINO-T [34]	O365,GoldG,RefC	✓	89.19	91.86	85.99	81.09	87.40	74.71	84.15	84.94
11	HyperLearner (Ours)	O365,GoldG	×	**50.66**	**60.87**	**44.66**	**59.29**	**62.29**	**45.43**	**67.02**	**67.44**
12	HyperLearner (Ours)	O365,GoldG,RefC	×	**77.89**	**76.92**	**72.99**	**67.54**	**75.55**	**57.54**	**77.00**	**76.79**
13	HyperLearner (Ours)	O365,GoldG,RefC	✓	**90.74**	**92.09**	85.46	**82.35**	84.70	72.64	82.53	82.39

Table 3. Comparison on RefCOCO/+/g benchmark. Metric: Top-1 accuracy.

Method	Backbone	Pre-training Data	Test AP$_{avg}$	
			zero-shot	full-shot
Detic-R [65]	RN50	LVIS, COCO, IN-21K	29.4	64.4
Detic-B [65]	Swin-B	LVIS, COCO, IN-21K	38.7	70.1
GLIP-T(A) [29]	Swin-T	O365	28.7	63.6
GLIP-T(B) [29]	Swin-T	O365	33.2	62.7
GLIP-T(C) [29]	Swin-T	O365, GoldG	44.4	63.9
Grounding-DINO-T [34]	Swin-T	O365, GoldG,Cap4M	44.9	-
HyperLearner (Ours)	Swin-T	O365	37.9	66.7
HyperLearner (Ours)	Swin-T	O365, GoldG	**45.2**	**68.9**

Table 4. Comparison on ODinW benchmark. Metric: mAP.

	Method	COCO	LVIS				ODinW
			AP	APr	APc	APf	
1	$\mathcal{L}_{Det}$	42.3	13.3	8.3	11.9	15.4	26.3
2	$\mathcal{L}_{baseline}$	46.8	23.6	22.7	24.3	24.6	38.6
3	$\mathcal{L}_{hyper}$	**48.4**	**31.3**	**30.7**	**32.6**	**30.3**	**45.2**

Table 5. Ablation study on learning objectives. Metric: mAP.
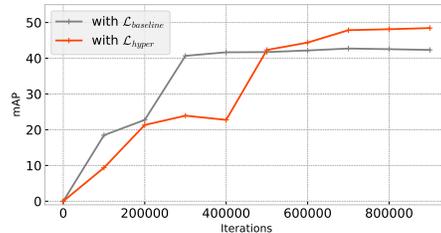

Figure 4. Comparison between baseline objective Eq. (15) and our objective Eq. (14) on COCO during training. Metric: mAP.

	Method	COCO	LVIS	ODinW
1	no synthetic captions	42.3	13.3	26.3
2	synthetic captions on $\mathcal{B}$	47.1	23.4	36.8
3	synthetic captions on $\mathcal{B} \cup \mathcal{G}$	47.6	23.7	38.6
4	synthetic captions on $\mathcal{B} \cup \mathcal{P}$	47.4	27.1	38.9
5	synthetic captions on $\mathcal{B} \cup \mathcal{G} \cup \mathcal{P}$	**48.4**	**31.3**	**45.2**

Table 6. Ablation study on region sampling for bootstrapping synthetic captions. Metric: mAP.

zero-shot testing scenario. It outperforms GLIP by +4.7% with pre-training solely on O365 and +0.8% when GoldG is added. While our model shows the commendable zero-shot transferability, it also excels in the full-shot setting, suggesting the pre-trained model possesses strong generalizablity.

Comparison on RefCOCO/+/g. Table 3 shows the results of localizing objects with referring expressions. When pre-trained on O365, GoldG or pre-trained on O365, GoldG, RefC, our model (in row 11, 12) surpasses Grounding DINO (in row 8, 9) and GLIP (in row 6). With further fine-tuning on RefC, our model (in row 13) performs competitively compared to Grounding DINO (in row 10). The competitive results in Table 3 can be attributed to learning with our bootstrapped captions, which often include descriptions of object attributes and their spatial relationships that also exist in the natural language expressions in RefCOCO/+/g.

4.3. Ablation study

We now ablate our approach and its components: hyperbolic learning objective (§3.3), region sampling for bootstrapping synthetic captions (§3.2), and cross-modal attention module for localization with free-form texts (§3.1).

Ablation study on learning objectives. Table 5 shows the results of different objectives: (1) standard open-world detection objective $\mathcal{L}_{Det}$ in Eq. (1), (2) vanilla contrastive learning objective $\mathcal{L}_{baseline}$ in Eq. (15), (3) hyperbolic learning objective $\mathcal{L}_{hyper}$ in Eq. (14). Our results show that our objective $\mathcal{L}_{hyper}$ outperforms other alternatives signifi-

cantly, consistently across different datasets. The improvement of $\mathcal{L}_{baseline}$ over $\mathcal{L}_{Det}$ shows the strength of learning with synthetic captions to boost open-world generalization; while the improvement of $\mathcal{L}_{hyper}$ over $\mathcal{L}_{baseline}, \mathcal{L}_{Det}$ further shows the benefit of aligning visual and caption embeddings in hierarchy with hyperbolic geometry (Figure 3).

Figure 4 compares the baseline objective $\mathcal{L}_{baseline}$ and our objective $\mathcal{L}_{hyper}$ during training. When using vanilla contrastive learning ($\mathcal{L}_{baseline}$), the performance saturates quickly. In contrast, our HyperLearner converges slower but consistently improves its performance, suggesting its stronger capability of learning from synthetic captions.

Ablation study on region sampling. Table 6 shows the results of bootstrapping richer synthetic captions on differ-

Figure 5. Qualitative results on open-world detection. Given any new class labels (row 1) or free-form texts (row 2) that specify objects with attribute, action, interaction, and spatial relationship, our model can detect and localize the objects in images.

|	Method	RefCOCO			COCO
		val	testA	testB	
1	*w/o* cross-modal attention	60.10	70.96	48.58	**48.7**
2	*w* cross-modal attention	**77.89**	**76.92**	**72.99**	48.4

Table 7. Ablation study on cross-modal attention.

ent regions: basic groundtruth bboxes $\mathcal{B}$, proposal sampling bboxes $\mathcal{P}$, grid sampling bboxes $\mathcal{G}$ (§3.2, Figure 2). As shown in row 1 and 2, utilizing captions on $\mathcal{B}$ improves the performance substantially, with a ≈10% increase on LVIS and ODinW. Moreover, the object-centric proposal sampling $\mathcal{P}$ (row 4) can target on novel objects, hence improving the detection on LVIS and ODinW, where performance are more dominated by rare novel objects. Finally, combining all sampled regions (row 5) yields the best overall performance, suggesting that using a diverse set of regions leads to richer semantics information and higher robustness.

Ablation study on cross-modal attention. Table 7 shows the results of using cross-modal attention to fuse spatial visual information and language semantics (Eq. (3), (4), (5) in §3.1). Our results indicate that the cross-modal attention module improves our model performance significantly on RefCOCO benchmark and yields similar performance on COCO benchmark. These results prove the crucial effectiveness of leveraging spatial and language information for the task of localization with free-form texts, as evidenced by the substantial boosts of 17.79%, 5.96%, 24.31% in Top-1 accuracy on the RefCOCO val, testA, testB respectively.

4.4. Qualitative analysis

Qualitative results. Figure 5 shows our qualitative results. In row 1, given new labels, our model can detect these new objects, even though these objects are small-scale (examples 2 and 3) and rare (example 4). In row 2, given any free-form texts that specify novel objects with fine-grained details, our model can localize these objects precisely, such as localizing the object of an attribute (*pink*), performing
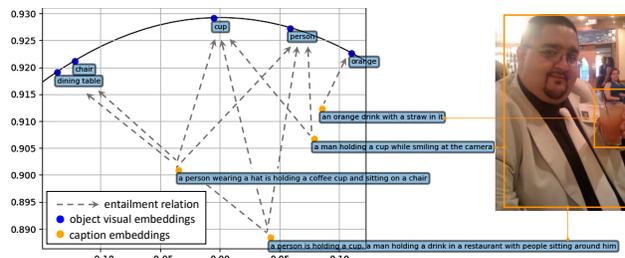
Figure 6. Visualization of our object and caption embeddings.

an action (*standing*), interacting with other object (*holding football*), or in a certain spatial location (*in the middle*). These examples indicate the strong generalizability of our model on recognizing new concepts in open world.

Visualization of embeddings. Figure 6 shows our object visual embeddings, caption embeddings with UMAP [38], which projects the learned embeddings into a 2D space. We can find that the visualized embeddings present a tree-like entailment relation; this is in line with our motivation of imposing the hierarchy of *'caption entails object'* in our learning objective (Figure 3). Interestingly, we can also find the hierarchy of *'longer caption entails shorter caption'*, which suggests the intrinsic hierarchy inside the caption data.

5. Conclusion

We present a novel hyperbolic vision-language learning approach to learn from synthetic captions effectively. Our approach is well designed to leverage the open-world knowledge from pre-trained VLMs, and specially mitigate the hallucination issue in synthetic captions to boost the open-world generalizability in detection. Our comprehensive experiments on multiple benchmark datasets show the competitive performance of our approach as compared to the state-of-the-art. We analyze our model design rationales with insightful ablation study and qualitative analysis. Our work also paved a strong foundation on leveraging synthetic captions with hyperbolic learning for other vision tasks.

Supplementary Material

A. Implementation details

In Table A, we present the learning schedules when fine-tuning on different pre-training datasets. Note that when adding GoldG, or RefC in row 2, 3, our model is initialized from the checkpoints in row 1, 2 respectively to continue model training. In Table B, we present the learning schedules when fine-tuning on different downstream datasets. During both pre-training and fine-tuning, the learning rate is decayed by 0.1 at 1/3 and 2/3 of the training progress.

Model	Dataset(s)	Iterations	batch size	learning rate
	O365	900,000	64	0.0002
HyperLearner	O365,GoldG	+500,000	64	0.0002
	O365,GoldG,RefC	+500,000	64	0.0002

Table A. Learning schedules on different pre-training datasets.

Model	Dataset(s)	Iterations	batch size	learning rate
	COCO	100,000	64	0.0001
HyperLearner	ODinW	50,000	64	0.0001
	RefCOCO	100,000	32	0.0001

Table B. Learning schedules on different downstream datasets.

B. Computation cost of hyperbolic learning

To mitigate the high computational cost of hyperbolic learning, we make the following design choices. First, we choose to only lift the last-layer Euclidean features into hyperbolic space, instead of using a hyperbolic neural network. This approach circumvents computationally intensive operations, such as Frechet mean calculations, as noted in [16].

Second, we adopt the Lorentz model as [36], rather than the Poincaré ball model employed in [8]. The Poincaré ball model is operationally intensive due to its computationally heavy Gyrovector operations, which necessitate a series of intermediate calculations. In contemporary deep learning frameworks, these calculations substantially increase the overhead of computation graph, as described in [13]. In contrast, the Lorentz model offers a more computationally efficient alternative for hyperbolic geometry, which employs four-vector operations (Euclidean axes plus a time dimension), thus learning only one additional dimension. Notably, the time dimension can be readily determined based on the curvature of the hyperboloid, as detailed in [8].

Third, thanks to the two design choices made above, the only additional computational costs are hyperbolic functions, including "sinh" and "cosh". Fortunately, PyTorch efficiently optimizes these operations, particularly through techniques such as operation fusion. This technique combines multiple pointwise operations – including arithmetic and trigonometric functions – into a single kernel, thus further enhancing computational efficiency, as detailed in [8].

C. Discussion on evaluation benchmarks

Open-vocabulary detection evaluation benchmarks. It is worth noting that existing works in open-vocabulary detection (OVD) [4, 15, 28, 50, 51, 59, 60, 64] consider the evaluation setting where only *novel* (*unseen*) object classes are presented in the evaluation benchmarks. Specifically, the OVCOCO and OVLVIS benchmarks are built by splitting the class labels into two sets: seen and unseen. The model is trained on the seen classes and tested on the unseen classes.

Open-world detection evaluation benchmarks. In contrast, our open-world detection task consider the evaluation setting where any object classes (both seen and unseen) are presented in the evaluation benchmarks, which aligns with detection in the open world. Moreover, these object classes could be described by keywords (class labels) or free-form texts. Our evaluation setting also poses a unique challenge of understanding novel concepts in free-form texts.

D. Additional results

Additional quantitative results on ODinW. Table C shows the additional results of 13 individual datasets on ODinW benchmark of our proposed approach, when using different evaluation modes (zero-shot or fine-tune/full-shot) and different pre-training data. Results show the benefit of using more data for pre-training, and further fine-tuning on different target datasets to improve generalization.

Additional qualitative results. Figure A presents more qualitative examples using our model to localize objects described using free-form texts to specify certain object attribute, action/interaction, and spatial relationships. Examples illustrate the strong generalizability of our model in understanding and recognizing novel concepts in open world.

E. Discussion on quality of synthetic captions

Generally, our VLM, BLIP2 6.7b, generates synthetic captions that accurately describe the objects in the image, as indicated by [28]. Furthermore, our experimental result reveals that the synthetic captions generated by VLM can significantly improve the performance of our model across various open-world object detection benchmarks. These observations qualitatively demonstrates the quality of synthetic captions for training open-world object detection models. In this section, we carry out a quantitative study for the quality of synthetic captions generated by VLMs, by determining the ratio of noisy or hallucinated objects in generated captions.

To quantify the noise level in our synthetic captions (i.e., % incorrect objects described), we adopt the Caption Hallucination Assessment with Image Relevance (CHAIR) metric from [31, 44] as our noise-level measurement for syn-

Model	Evaluation Mode	Data	AerialDrone	Aquarium	Rabbits	EgoHands	Mushrooms	Packages	PascalVOC	pistols	pothole	Raccoon	Shellfish	Thermal	Vehicles	mAP
HyperLearner	zero-shot	O365	16.63	27.35	72.53	12.47	41.08	63.68	55.88	23.03	15.31	42.68	21.50	40.51	59.92	36.05
	fine-tune	O365	41.53	50.72	75.59	68.41	83.45	72.34	84.39	99.77	52.08	61.00	43.09	75.35	59.42	67.31
	zero-shot	O365, GoldG	27.25	35.34	74.38	18.12	58.50	68.95	64.12	27.03	17.06	56.01	29.36	60.72	60.24	44.74
	fine-tune	O365, GoldG	42.48	52.10	76.34	74.66	90.43	76.53	87.94	99.77	53.00	63.58	39.52	78.52	61.25	69.57

Table C. Detailed results of zero-shot and fine-tune transfer on 13 individual datasets on ODinW. Metric: mAP over 13 datasets.
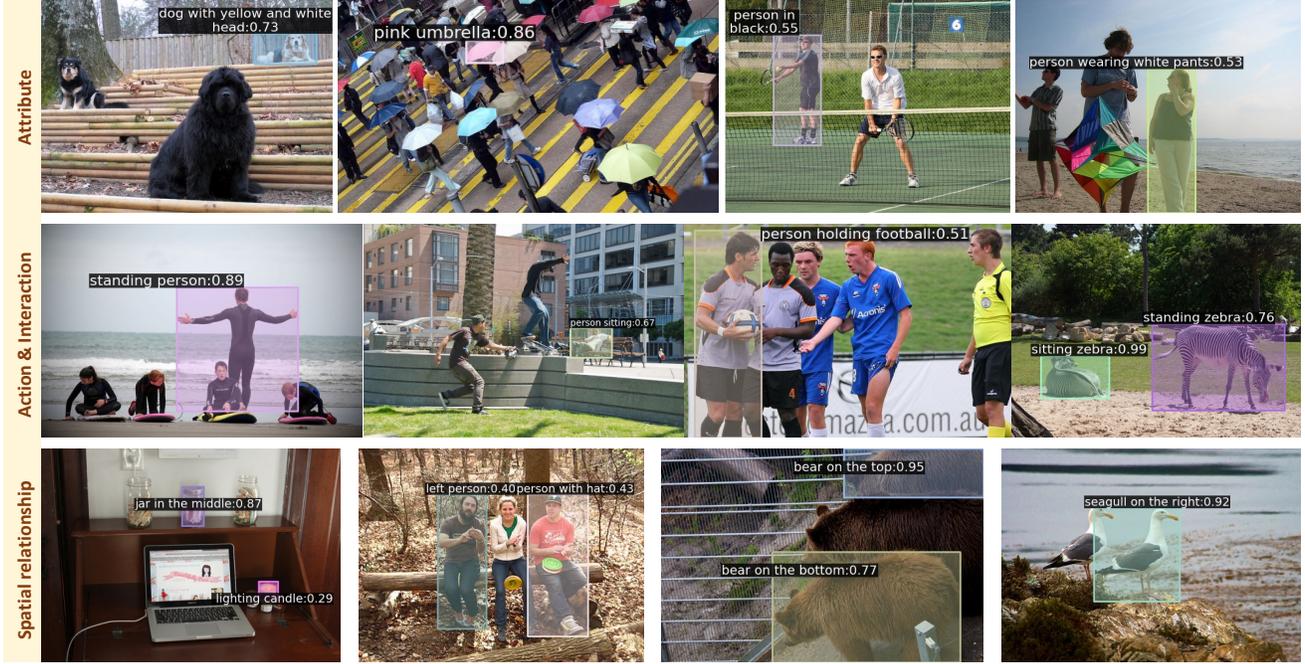

Figure A. More qualitative results on open-world detection using free-form texts, where the free-form texts specify objects with attribute, action/interaction, and spatial relationships.

thetic captions. It can be formulated as the percentage of incorrectly described objects,

$$\text{noise \%} = \frac{1}{m} \sum_{i=1}^{m} \frac{N_{\text{incorrect objects}}}{N_{\text{total objects}}} \times 100\% \qquad (16)$$

,where $N_{\text{incorrect objects}}$ denotes the number of hallucinated objects in the caption, $N_{\text{total objects}}$ denotes the number of total objects in the caption, and m is the total number of synthetic captions collected from the dataset. To obtain $N_{\text{incorrect objects}}$ and $N_{\text{total objects}}$ in (Eq. (16)), we detect the synonyms of object classes within the synthetic captions using WordNet synsets [39]. If an object or its synonym is mentioned in the caption, but the detection ground truth does not include this object, it is classified as an incorrect object. Our result indicates a noise level of 16.3% on COCO 2017 test set, using BLIP2 6.7b. Despite the presence of noise and hallucinated contents in the synthetic captions, we have proposed a hyperbolic learning strategy to mitigate the issue of training with noisy synthetic captions. The hyperbolic learning loss learns visual and caption embeddings hierarchically, imposing an entailment relationship between captions and region embeddings, rather than directly aligning them. Moreover, we suggest that employing an enhanced pre-trained VLM for captioning could further

mitigate the noise level.

F. Future work

Building upon our current work in open-world object detection, there are several promising avenues for future exploration. First, it is essential to develop a new benchmark that evaluates open-world object detection using both keywords and free-form text, overcoming the limitations of current benchmarks that merely focus on either keywords (e.g., COCO, LVIS, ODinW) or free-form texts (e.g., RefCOCO). Second, training image captioning models that generate object-centric synthetic captions presents an intriguing research direction to provide better open-world knowledge in object detection. Third, integrating various data modalities or information sources to augment the model's detection capabilities is also an interesting direction, such as providing a sketch of a target object to localize in the open world.

References

[1] Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. End-to-end object detection with transformers. In *ECCV*, 2020. 2

[2] Ines Chami, Adva Wolf, Da-Cheng Juan, Frederic Sala, Sujith Ravi, and Christopher Ré. Low-dimensional hyperbolic knowledge graph embeddings. *arXiv preprint arXiv:2005.00545*, 2020. 2

[3] Yanbei Chen, Manchen Wang, Abhay Mittal, Zhenlin Xu, Paolo Favaro, Joseph Tighe, and Davide Modolo. Scaledet: A scalable multi-dataset object detector. In *CVPR*, 2023. 2

[4] Han-Cheol Cho, Won Young Jhoo, Wooyoung Kang, and Byungseok Roh. Open-vocabulary object detection using pseudo caption labels. *arXiv preprint arXiv:2303.13040*, 2023. 2, 9

[5] Shuyang Dai, Zhe Gan, Yu Cheng, Chenyang Tao, Lawrence Carin, and Jingjing Liu. Apo-vae: Text generation in hyperbolic space. *arXiv preprint arXiv:2005.00054*, 2020. 2

[6] Xiyang Dai, Yinpeng Chen, Bin Xiao, Dongdong Chen, Mengchen Liu, Lu Yuan, and Lei Zhang. Dynamic head: Unifying object detection heads with attentions. In *CVPR*, 2021. 6

[7] Karan Desai and Justin Johnson. Virtex: Learning visual representations from textual annotations. In *CVPR*, 2021. 2

[8] Karan Desai, Maximilian Nickel, Tanmay Rajpurohit, Justin Johnson, and Shanmukha Ramakrishna Vedantam. Hyperbolic image-text representations. In *ICML*, 2023. 2, 4, 5, 9

[9] Bhuwan Dhingra, Christopher J Shallue, Mohammad Norouzi, Andrew M Dai, and George E Dahl. Embedding text in hyperbolic spaces. *arXiv preprint arXiv:1806.04313*, 2018. 2

[10] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020. 3

[11] Sivan Doveh, Assaf Arbelle, Sivan Harary, Amit Alfassy, Roei Herzig, Donghyun Kim, Raja Giryes, Rogerio Feris, Rameswar Panda, Shimon Ullman, et al. Dense and aligned captions (dac) promote compositional reasoning in vl models. In *NeurIPS*, 2023. 2

[12] Octavian Ganea, Gary Bécigneul, and Thomas Hofmann. Hyperbolic entailment cones for learning hierarchical embeddings. In *ICML*, 2018. 2

[13] Songwei Ge, Shlok Mishra, Simon Kornblith, Chun-Liang Li, and David Jacobs. Hyperbolic contrastive learning for visual representations beyond objects. In *CVPR*, 2023. 2, 4, 9

[14] Ross Girshick. Fast r-cnn. In *ICCV*, 2015. 2, 6

[15] Xiuye Gu, Tsung-Yi Lin, Weicheng Kuo, and Yin Cui. Open-vocabulary object detection via vision and language knowledge distillation. In *ICLR*, 2021. 1, 2, 9

[16] Yunhui Guo, Xudong Wang, Yubei Chen, and Stella X Yu. Clipped hyperbolic classifiers are super-hyperbolic classifiers. In *CVPR*, 2022. 9

[17] Agrim Gupta, Piotr Dollar, and Ross Girshick. Lvis: A dataset for large vocabulary instance segmentation. In *CVPR*, 2019. 5

[18] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. Mask r-cnn. In *ICCV*, 2017. 3

[19] Chao Jia, Yinfei Yang, Ye Xia, Yi-Ting Chen, Zarana Parekh, Hieu Pham, Quoc Le, Yun-Hsuan Sung, Zhen Li, and Tom Duerig. Scaling up visual and vision-language representation learning with noisy text supervision. In *ICML*, 2021. 2

[20] Aishwarya Kamath, Mannat Singh, Yann LeCun, Gabriel Synnaeve, Ishan Misra, and Nicolas Carion. Mdetr-modulated detection for end-to-end multi-modal understanding. In *ICCV*, 2021. 5, 6, 7

[21] Sahar Kazemzadeh, Vicente Ordonez, Mark Matten, and Tamara Berg. Referitgame: Referring to objects in photographs of natural scenes. In *EMNLP*, 2014. 6

[22] Valentin Khrulkov, Leyla Mirvakhabova, Evgeniya Ustinova, Ivan Oseledets, and Victor Lempitsky. Hyperbolic image embeddings. In *CVPR*, 2020. 2

[23] Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, Li-Jia Li, David A Shamma, et al. Visual genome: Connecting language and vision using crowdsourced dense image annotations. *IJCV*, 2017. 3

[24] Weicheng Kuo, Fred Bertsch, Wei Li, AJ Piergiovanni, Mohammad Saffar, and Anelia Angelova. Findit: Generalized localization with natural language queries. In *ECCV*, 2022. 1

[25] Marc Law, Renjie Liao, Jake Snell, and Richard Zemel. Lorentzian distance learning for hyperbolic representations. In *ICML*, 2019. 4, 5

[26] Matt Le, Stephen Roller, Laetitia Papaxanthos, Douwe Kiela, and Maximilian Nickel. Inferring concept hierarchies from text corpora via hyperbolic embeddings. *arXiv preprint arXiv:1902.00913*, 2019. 5

[27] Gen Li, Nan Duan, Yuejian Fang, Ming Gong, and Daxin Jiang. Unicoder-vl: A universal encoder for vision and language by cross-modal pre-training. In *Proceedings of the AAAI conference on artificial intelligence*, volume 34, pages 11336–11344, 2020. 3

[28] Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. *arXiv preprint arXiv:2301.12597*, 2023. 2, 3, 5, 9

[29] Liunian Harold Li, Pengchuan Zhang, Haotian Zhang, Jianwei Yang, Chunyuan Li, Yiwu Zhong, Lijuan Wang, Lu Yuan, Lei Zhang, Jenq-Neng Hwang, et al. Grounded language-image pre-training. In *CVPR*, 2022. 1, 2, 3, 5, 6, 7

[30] Muchen Li and Leonid Sigal. Referring transformer: A one-step approach to multi-task visual grounding. In *NeurIPS*, 2021. 7

[31] Yifan Li, Yifan Du, Kun Zhou, Jinpeng Wang, Wayne Xin Zhao, and Ji-Rong Wen. Evaluating object hallucination in large vision-language models. *arXiv preprint arXiv:2305.10355*, 2023. 9

[32] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *ECCV*, 2014. 1, 5

[33] Shilong Liu, Shijia Huang, Feng Li, Hao Zhang, Yaoyuan

Liang, Hang Su, Jun Zhu, and Lei Zhang. Dq-detr: Dual query detection transformer for phrase extraction and grounding. In *AAAI*, 2023. 7

[34] Shilong Liu, Zhaoyang Zeng, Tianhe Ren, Feng Li, Hao Zhang, Jie Yang, Chunyuan Li, Jianwei Yang, Hang Su, Jun Zhu, et al. Grounding dino: Marrying dino with grounded pre-training for open-set object detection. *arXiv preprint arXiv:2303.05499*, 2023. 2, 6, 7

[35] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *ICCV*, 2021. 5

[36] Aaron Lou, Isay Katsman, Qingxuan Jiang, Serge Belongie, Ser-Nam Lim, and Christopher De Sa. Differentiating through the fréchet mean. In *ICML*, 2020. 9

[37] Junhua Mao, Jonathan Huang, Alexander Toshev, Oana Camburu, Alan L Yuille, and Kevin Murphy. Generation and comprehension of unambiguous object descriptions. In *CVPR*, 2016. 6, 7

[38] Leland McInnes, John Healy, and James Melville. Umap: Uniform manifold approximation and projection for dimension reduction. *arXiv preprint arXiv:1802.03426*, 2018. 8

[39] George A Miller. Wordnet: a lexical database for english. *Communications of the ACM*, 38(11):39–41, 1995. 10

[40] Maximillian Nickel and Douwe Kiela. Learning continuous hierarchies in the lorentz model of hyperbolic geometry. In *ICML*, 2018. 5

[41] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *ICML*, 2021. 2, 3, 4, 5

[42] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. In *NeurIPS*, 2015. 3, 6

[43] Tal Ridnik, Emanuel Ben-Baruch, Asaf Noy, and Lihi Zelnik-Manor. Imagenet-21k pretraining for the masses. *arXiv preprint arXiv:2104.10972*, 2021. 3

[44] Anna Rohrbach, Lisa Anne Hendricks, Kaylee Burns, Trevor Darrell, and Kate Saenko. Object hallucination in image captioning. *arXiv preprint arXiv:1809.02156*, 2018. 2, 4, 9

[45] Shuai Shao, Zeming Li, Tianyuan Zhang, Chao Peng, Gang Yu, Xiangyu Zhang, Jing Li, and Jian Sun. Objects365: A large-scale, high-quality dataset for object detection. In *ICCV*, 2019. 5

[46] Piyush Sharma, Nan Ding, Sebastian Goodman, and Radu Soricut. Conceptual captions: A cleaned, hypernymed, image alt-text dataset for automatic image captioning. In *ACL*, 2018. 1, 3

[47] Zhi Tian, Chunhua Shen, Hao Chen, and Tong He. Fcos: Fully convolutional one-stage object detection. In *ICCV*, 2019. 3

[48] Michael Tschannen, Manoj Kumar, Andreas Steiner, Xiaohua Zhai, Neil Houlsby, and Lucas Beyer. Image captioners are scalable vision learners too. *arXiv preprint arXiv:2306.07915*, 2023. 2

[49] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *NeurIPS*, 2017. 3

[50] Size Wu, Wenwei Zhang, Sheng Jin, Wentao Liu, and Chen Change Loy. Aligning bag of regions for open-vocabulary object detection. In *CVPR*, 2023. 1, 2, 9

[51] Xiaoshi Wu, Feng Zhu, Rui Zhao, and Hongsheng Li. Cora: Adapting clip for open-vocabulary detection with region prompting and anchor pre-matching. In *CVPR*, 2023. 2, 9

[52] Jianwei Yang, Chunyuan Li, Pengchuan Zhang, Bin Xiao, Ce Liu, Lu Yuan, and Jianfeng Gao. Unified contrastive learning in image-text-label space. In *CVPR*, 2022. 2

[53] Kaicheng Yang, Jiankang Deng, Xiang An, Jiawei Li, Ziyong Feng, Jia Guo, Jing Yang, and Tongliang Liu. Alip: Adaptive language-image pre-training with synthetic caption. In *ICCV*, 2023. 2

[54] Lewei Yao, Jianhua Han, Youpeng Wen, Xiaodan Liang, Dan Xu, Wei Zhang, Zhenguo Li, Chunjing Xu, and Hang Xu. Detclip: Dictionary-enriched visual-concept paralleled pretraining for open-world detection. In *NeurIPS*, 2022. 6

[55] Lewei Yao, Runhui Huang, Lu Hou, Guansong Lu, Minzhe Niu, Hang Xu, Xiaodan Liang, Zhenguo Li, Xin Jiang, and Chunjing Xu. Filip: Fine-grained interactive language-image pre-training. In *ICLR*, 2021. 2

[56] Jiahui Yu, Zirui Wang, Vijay Vasudevan, Legg Yeung, Mojtaba Seyedhosseini, and Yonghui Wu. Coca: Contrastive captioners are image-text foundation models. *arXiv preprint arXiv:2205.01917*, 2022. 2

[57] Licheng Yu, Zhe Lin, Xiaohui Shen, Jimei Yang, Xin Lu, Mohit Bansal, and Tamara L Berg. Mattnet: Modular attention network for referring expression comprehension. In *CVPR*, 2018. 6, 7

[58] Licheng Yu, Patrick Poirson, Shan Yang, Alexander C Berg, and Tamara L Berg. Modeling context in referring expressions. In *ECCV*, 2016. 6

[59] Yuhang Zang, Wei Li, Kaiyang Zhou, Chen Huang, and Chen Change Loy. Open-vocabulary detr with conditional matching. In *ECCV*, 2022. 2, 9

[60] Alireza Zareian, Kevin Dela Rosa, Derek Hao Hu, and Shih-Fu Chang. Open-vocabulary object detection using captions. In *CVPR*, 2021. 1, 9

[61] Xiaohua Zhai, Xiao Wang, Basil Mustafa, Andreas Steiner, Daniel Keysers, Alexander Kolesnikov, and Lucas Beyer. Lit: Zero-shot transfer with locked-image text tuning. In *CVPR*, 2022. 2

[62] Hao Zhang, Feng Li, Shilong Liu, Lei Zhang, Hang Su, Jun Zhu, Lionel M Ni, and Heung-Yeung Shum. Dino: Detr with improved denoising anchor boxes for end-to-end object detection. *arXiv preprint arXiv:2203.03605*, 2022. 6

[63] Haotian* Zhang, Pengchuan* Zhang, Xiaowei Hu, Yen-Chun Chen, Liunian Harold Li, Xiyang Dai, Lijuan Wang, Lu Yuan, Jenq-Neng Hwang, and Jianfeng Gao. Glipv2: Unifying localization and vision-language understanding. In *NeurIPS*, 2022. 1, 2

[64] Yiwu Zhong, Jianwei Yang, Pengchuan Zhang, Chunyuan Li, Noel Codella, Liunian Harold Li, Luowei Zhou, Xiyang Dai, Lu Yuan, Yin Li, et al. Regionclip: Region-based language-image pretraining. In *CVPR*, 2022. 2, 9

[65] Xingyi Zhou, Rohit Girdhar, Armand Joulin, Phillip Krähenbühl, and Ishan Misra. Detecting twenty-thousand classes using image-level supervision. In *ECCV*, 2022. 1, 2, 3, 5, 6, 7

[66] Xingyi Zhou, Vladlen Koltun, and Philipp Krähenbühl. Probabilistic two-stage detection. *arXiv preprint arXiv:2103.07461*, 2021. 3, 6

[67] Xingyi Zhou, Dequan Wang, and Philipp Krähenbühl. Objects as points. *arXiv preprint arXiv:1904.07850*, 2019. 3

[68] Xizhou Zhu, Weijie Su, Lewei Lu, Bin Li, Xiaogang Wang, and Jifeng Dai. Deformable detr: Deformable transformers for end-to-end object detection. In *ICLR*, 2020. 6