# SHiNe: Semantic Hierarchy Nexus for Open-vocabulary Object Detection

Mingxuan Liu[1,2*] Tyler L. Hayes[2] Elisa Ricci[1,3] Gabriela Csurka[2] Riccardo Volpi[2]
[1] University of Trento [2] NAVER LABS Europe [3] Fondazione Bruno Kessler

## Abstract

*Open-vocabulary object detection (OvOD) has transformed detection into a language-guided task, empowering users to freely define their class vocabularies of interest during inference. However, our initial investigation indicates that existing OvOD detectors exhibit significant variability when dealing with vocabularies across various semantic granularities, posing a concern for real-world deployment. To this end, we introduce **S**emantic **Hi**erarchy **Ne**xus (**SHiNe**), a novel classifier that uses semantic knowledge from class hierarchies. It runs offline in three steps: i) it retrieves relevant super-/sub-categories from a hierarchy for each target class; ii) it integrates these categories into hierarchy-aware sentences; iii) it fuses these sentence embeddings to generate the nexus classifier vector. Our evaluation on various detection benchmarks demonstrates that SHiNe enhances robustness across diverse vocabulary granularities, achieving up to +31.9% mAP50 with ground truth hierarchies, while retaining improvements using hierarchies generated by large language models. Moreover, when applied to open-vocabulary classification on ImageNet-1k, SHiNe improves the CLIP zero-shot baseline by +2.8% accuracy. SHiNe is training-free and can be seamlessly integrated with any off-the-shelf OvOD detector, without incurring additional computational overhead during inference. The code is open source.*

*A complicated series of connections between different things.*

Definition of *Nexus*, **Oxford Dictionary**

## 1. Introduction

Open-vocabulary object detection (OvOD) [18, 59, 65, 73] transforms the object detection task into a language-guided matching problem between visual regions and class names. Leveraging weak supervisory signals and a pre-aligned vision-language space from Vision-Language Models (VLMs) [22, 42], OvOD methods [18, 29, 65, 72, 73] extend the ability of models to localize and categorize objects beyond the trained categories. Under the OvOD paradigm, target object classes are described using text prompts like `"a {Class Name}"`, rather than class indices. By altering the `"{Class Name}"`, OvOD methods enable users to *freely* define their own Classes of Interest (CoIs) using natural language. This allows new classes of interest to be detected without the need for model re-training.

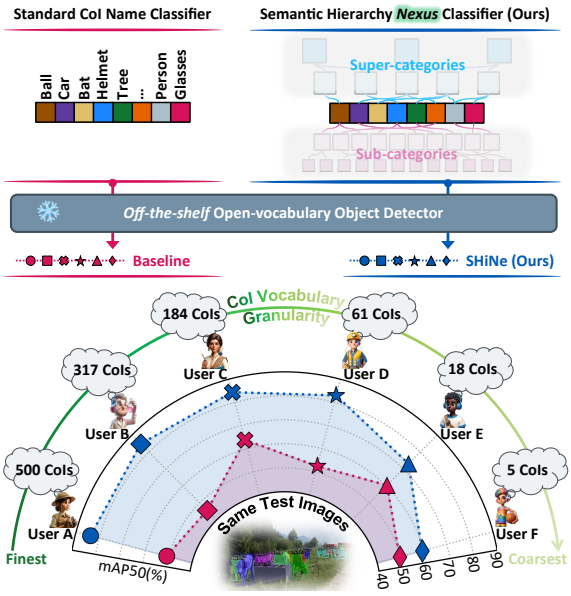


Figure 1. **(Top)** Classifier comparison for open-vocabulary object detectors: **(Left)** standard methods use solely class names in the vocabulary specified by the user to extract text embeddings; **(Right)** our proposed SHiNe fuses information from super-/sub-categories into *nexus* points to generate hierarchy-aware representations. **(Bottom)** Open-vocabulary detection performance at different levels of vocabulary granularity specified by users: A standard Baseline under-performs and presents significant variability; SHiNe allows for improved and more uniform performance across various vocabularies. Results are on the iNatLoc [6] dataset.

Yet, recent studies for open-vocabulary classification [14, 38, 40] highlight a key challenge: open-vocabulary methods are sensitive to the choice of vocabulary. For instance, Parashar *et al.* [40] enhanced CLIP's zero-shot performance by substituting scientific CoI names, like `"Rosa"`, with common English names, such as `"Rose"`. Recent OvOD models have improved performance by better aligning object features with the VLM semantic space [26, 60]. However, a pivotal question remains: *Are off-the-shelf*

*Correspondence to: mingxuan.liu@unitn.it

*OvOD detectors truly capable of handling an open vocabulary **across various semantic granularities**?*

In practical scenarios, Classes of Interest (CoIs) are in the eyes of the beholder. For example, consider a region crop of a `"Dog"`: one user may be interested in the specific breed (*e.g.*, `"Labrador"`), while another might only be concerned about whether it is an `"Animal"`. Thus, the CoI is defined at varying levels of semantic granularity. Ideally, since these CoIs refer to the same visual region, the performance of an OvOD detector should be consistent across different granularities. However, our initial experiments (illustrated in Fig. 1) reveal that the performance of an OvOD detector [72] (see **Baseline**) fluctuates based on the vocabulary granularity. This inconsistency in performance across granularities presents a significant concern for deploying *off-the-shelf* OvOD models in real-world contexts, especially in safety-critical [25] areas like autonomous driving [34].

Although the same physical object, a `"Labrador"`, can be classified at varying levels of granularity, the inherent *fact* that a `"Labrador is a dog, which is an animal"` remains *constant*. This knowledge is readily available from a semantic hierarchy. Guided by this rationale, we aim to enhance the robustness of existing OvOD detectors to vocabularies specified at any granularity by leveraging knowledge inherent in semantic hierarchies. Recent research in open-vocabulary classification [14, 38] has explored using super-/sub-categories of CoIs from hierarchies to improve accuracy. However, these methods involve searching through sub-categories or both super-/sub-categories at inference time, leading to additional computational overhead and limiting their use in detection tasks.

We introduce the **S**emantic **Hi**erarchy ***Nexus*** (**SHiNe**), a novel classifier designed to enhance the robustness of OvOD to diverse vocabulary granularities. SHiNe is *training-free*, and ensures that the inference procedure is *linear* in complexity relative to the number of CoIs. SHiNe first retrieves relevant super(abstract)-/sub(specific)-categories from a semantic hierarchy for each CoI in a vocabulary. It then uses an **Is-A** connector to integrate these categories into hierarchy-aware sentences, while *explicitly* modeling their internal relationships. Lastly, it fuses these text embeddings into a vector, termed *nexus*, using an aggregator (*e.g.*, the mean operation) to form a classifier weight for the target CoI. SHiNe can be directly integrated with any *off-the-shelf* VLM-based OvOD detector. As shown in Fig. 1, SHiNe consistently improves performance across a range of CoI vocabulary granularities, while narrowing performance gaps at different granularities.

We evaluate SHiNe on various detection datasets [6, 10], that cover a broad range of label vocabulary granularities. This includes scenarios with readily available hierarchies and cases *without* them. In the latter, we utilize large language models [39] to generate a synthetic [38] three-level hierarchy for SHiNe. Our results demonstrate that SHiNe significantly and consistently improves the performance and robustness of baseline detectors, and showcase its generalizability to other *off-the-shelf* OvOD detectors. Additionally, we extend SHiNe to open-vocabulary classification and further validate its effectiveness by comparing it with two state-of-the-art methods [14, 38] on the ImageNet-1k [7] dataset. The key contributions of this work are:

- We show that the performance of existing OvOD detectors varies across vocabulary granularities. This highlights the need for enhanced robustness to arbitrary granularities, especially for real-world applications.
- We introduce SHiNe, a novel classifier that improves the robustness of OvOD models to various vocabulary granularities using semantic knowledge from hierarchies. SHiNe is *training-free* and compatible with existing and generated hierarchies. It can be seamlessly integrated into any OvOD detector *without* computational overhead.
- We demonstrate that SHiNe consistently enhances the performance of OvOD detectors across various vocabulary granularities on iNatLoc [6] and FSOD [10], with gains of up to **+31.9** points in mAP50. On open-vocabulary classification, SHiNe improves the CLIP [42] zero-shot baseline by up to **+2.8%** on ImageNet-1k [7].

## 2. Related Work

**Open-vocabulary object detection (OvOD)** [59, 73] is rapidly gaining traction due to its practical significance, allowing users to *freely* define their Classes of Interest (CoIs) during inference and facilitating the detection of newly specified objects in a zero-shot way. With the aid of weak supervisory signals, OvOD surpasses zero-shot detectors [54] by efficiently aligning visual region features with an embedding space that has been *pre-aligned* with image and text by contrastive vision-language models (VLMs) [22, 42]. This process is approached from either the vision or text side to bridge the gap between region-class and image-class alignments. To this end, methods based on region-aware training [61, 63–65], pseudo-labeling [1, 12, 68, 72], knowledge distillation [9, 18, 60], and transfer learning [26, 36, 67] are explored. In our study, we apply our method to pre-trained region-text aligned OvOD detectors, improving their performance and robustness to vocabularies of diverse granularities. Our method shares conceptual similarities with the work of Kaul *et al.* [23], where they develop a multi-modal classifier that merges a text-based classifier enriched with descriptors [35] from GPT-3 [4] and a vision classifier grounded in image exemplars. This classifier is then used to train an OvOD detector [72] with an extra *learnable* bias. In contrast, our proposed SHiNe is *training-free*, enabling effortless integration with any OvOD detector.

**Prompt engineering** [17] has been extensively studied as a technique to enhance VLMs [22, 42, 66]. *Prompt enrichment* methods [35, 40, 41, 45, 62] have focused on augmenting frozen VLM text classifiers by incorporating additional class descriptions sourced from large language models (LLMs) [4]. In contrast, our work explores the acquisition of useful semantic knowledge from a hierarchy. *Prompt tuning* methods [24, 43, 52, 57, 69, 70] introduced *learnable* token vectors into text prompts, which are fine-tuned on downstream tasks. In contrast, our proposed method is *training-free*. Our work is mostly related to two recent methods, CHiLS [38] and H-CLIP [14], that improve CLIP's [42] zero-shot classification performance by relying on a semantic hierarchy. CHiLS searches for higher logit score matches within the sub-categories, using the max score found to update the initial prediction. H-CLIP runs a combinatorial search over related super-/sub-categories prompt combinations for higher logit scores. However, both approaches incur additional computational overhead due to their *search-on-the-fly* mechanism during inference, constraining their use to classification tasks. In contrast, SHiNe operates offline and adds no overhead at inference, making it applicable to both classification and detection tasks.

**Semantic hierarchy** [6, 11, 55, 56] is a tree-like taxonomy [58] or a directed acyclic graph [47] that structures semantic concepts following an *asymmetric* and *transitive* "Is-A" relation [53]. Previous works have used such hierarchies to benefit various vision tasks [2, 3, 8, 13, 16, 37, 46]. Cole *et al.* [6] introduce the extensive iNatLoc dataset with a six-level hierarchy to enhance weakly supervised object localization, showing that appropriate label granularity can improve model training. Shin *et al.* [51] and Hamamci *et al.* [20] develop hierarchical architectures that incorporate multiple levels of a label hierarchy for training, enhancing multi-label object detection in remote sensing and dental X-ray images, respectively. Our work distinguishes itself from previous studies in two key ways: *i)* We focus on multi-modal models; *ii)* We improve OvOD detectors using label hierarchies as an external knowledge base, without requiring hierarchical annotations or any training. Furthermore, SHiNe does not rely on a ground-truth hierarchy and can work with an LLM-generated [39] hierarchy.

# 3. Method

Our objective is to improve the robustness of *off-the-shelf* open-vocabulary object detectors to diverse user-defined Classes of Interest (CoIs) with varying levels of semantic granularity. We first provide an introduction of open-vocabulary object detection (OvOD). Sec. 3.1 introduces our method of developing the **S**emantic **Hi**erarchy ***Nexus*** (SHiNe) based classifier for OvOD detectors to improve their vocabulary granularity robustness. Once established, the SHiNe classifier can be directly integrated with existing trained OvOD detectors and transferred to novel datasets in a zero-shot manner as discussed in Sec. 3.2.

**Problem formulation.** The objective of open-vocabulary object detection is to localize and classify novel object classes freely specified by the user within an image, without any retraining, in a zero-shot manner. Given an input image $\mathbf{I} \in \mathbb{R}^{3\times h\times w}$, OvOD localizes all foreground objects and classifies them by estimating a set of bounding box coordinates and class label pairs $\{\mathbf{b}_m, c_m\}_{m=1}^{M}$, with $\mathbf{b}_m \in \mathbb{R}^4$ and $c_m \in \mathcal{C}^{\text{test}}$, where $\mathcal{C}^{\text{test}}$ is the vocabulary set defined by the user at test time. To attain open-vocabulary capabilities, OvOD [31, 72, 73] uses a box-labeled dataset $\mathcal{D}^{\text{det}}$ with a limited vocabulary $\mathcal{C}^{\text{det}}$ and an auxiliary dataset $\mathcal{D}^{\text{weak}}$ as weak supervisory signals. $\mathcal{D}^{\text{weak}}$ features fewer detailed image-class or image-caption annotation pairs, yet it encompasses a broad vocabulary $\mathcal{C}^{\text{weak}}$ (*e.g.*, ImageNet-21k [7]), significantly expanding the detection lexicon.

**Open-vocabulary detector.** Predominant OvOD detectors, such as Detic [72] and VLDet [31], follow a two-stage pipeline. First, given an image, a learned region proposal network (RPN) yields a bag of $M$ region proposals by $\{\mathbf{z}_m\}_{m=1}^{M} = \Phi_{\text{RPN}}(\mathbf{I})$, where $\mathbf{z}_m \in \mathbb{R}^D$ is a $D$-dimensional region-of-interest (RoI) feature embedding. Then, for each proposed region, a learned bounding box regressor predicts the location coordinates by $\hat{\mathbf{b}}_m = \Phi_{\text{REG}}(\mathbf{z}_m)$. An open-vocabulary classifier estimates a set of classification scores $s_m(c, \mathbf{z}_m) = \langle \mathbf{w}_c, \mathbf{z}_m \rangle$ for each class, where $\mathbf{w}_c$ is a vector in the classifier $\mathbf{W} \in \mathbb{R}^{|\mathcal{C}^{\text{test}}|}$ and $\langle \cdot, \cdot \rangle$ is the cosine similarity function. $\mathbf{W}$ is the frozen text classifier, created by using a VLM text encoder (*e.g.*, CLIP [42]) to encode the names of CoIs in $\mathcal{C}^{\text{test}}$ specified by the user. The CoI that yields the highest score is assigned as the classification result. During training, OvOD detectors learn all model parameters except for the frozen text classifier. This allows them to achieve region-class alignment by leveraging the vision-language semantic space pre-aligned by VLMs for the open-vocabulary capability. Our work aims to improve existing pre-trained OvOD detectors, so we omit further details, and refer the reader to dedicated surveys [59, 73].

## 3.1. SHiNe: Semantic Hierarchy Nexus

Here, we describe SHiNe, our proposed semantic hierarchy *nexus*-based classifier for improving OvOD. As illustrated in Fig. 2(top), for each target CoI $c \in \mathcal{C}^{\text{test}}$ (*e.g.*, `"Bat"`) in the user-defined vocabulary, we construct a *nexus* point $\mathbf{n}_c \in \mathbb{R}^D$ by incorporating information from related super-/sub-categories derived from a semantic hierarchy $\mathcal{H}$. SHiNe is *training-free*. Upon constructing the *nexus* points for the entire vocabulary *offline*, the *nexus*-based classifier $\mathbf{N}$ is directly applied to an *off-the-shelf* OvOD detector for inference. This replaces the conventional CoI name-based classifier $\mathbf{W}$ with our hierarchy-aware SHiNe classifier. This enables the classification score $s_m(c, \mathbf{z}_m) = \langle \mathbf{n}_c, \mathbf{z}_m \rangle$

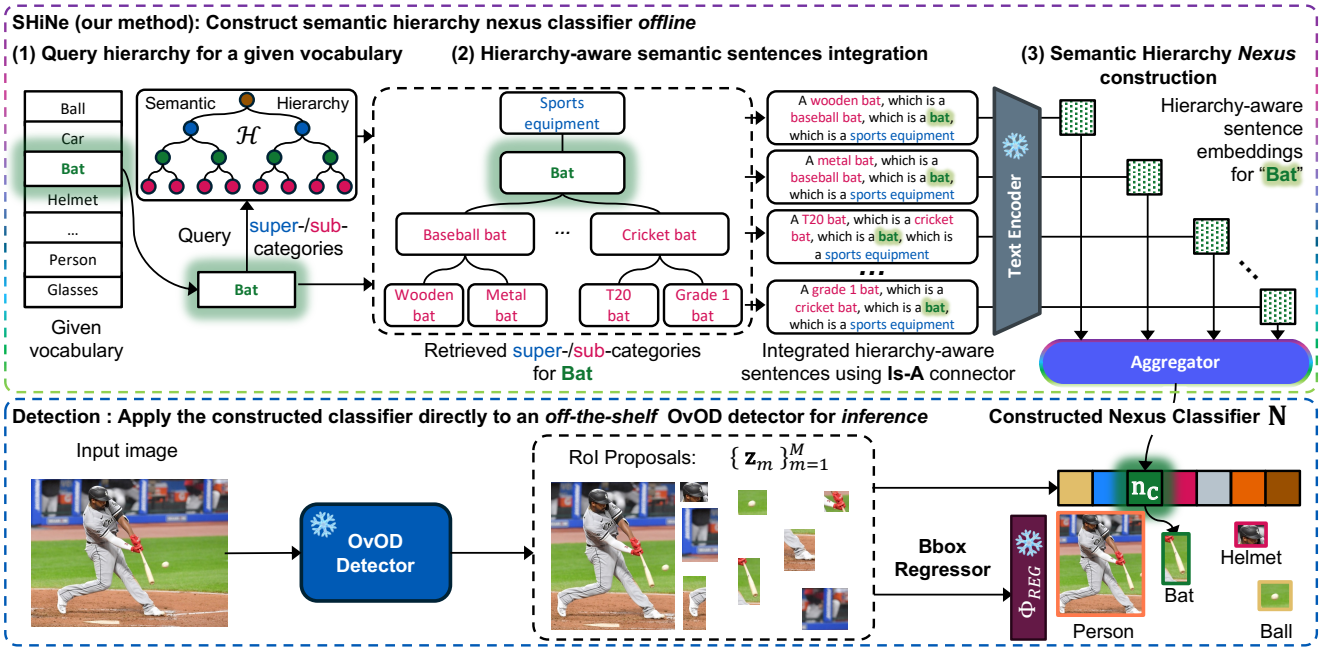


Figure 2. Overview of our method. **(Top)** SHiNe constructs the semantic hierarchy *nexus* classifier in three steps *offline*: (1) For each target class (*e.g.*, "Bat" in green) in the given vocabulary, we query the associated super-(in blue)/sub-(in pink) categories from a semantic hierarchy. (2) These retrieved categories along with their interrelationships are integrated into a set of hierarchy-aware sentences using our proposed **Is-A** connector. (3) These sentences are then encoded by a frozen VLM text encoder (*e.g.*, CLIP [42]) and subsequently fused using an aggregator (*e.g.*, mean-aggregator) to form a *nexus* classifier vector for the target class. **(Bottom)**: The constructed classifier is directly applied to an *off-the-shelf* OvOD detector for inference, enhancing its robustness across various levels of vocabulary granularity.

to be high when the proposed region closely aligns with the semantic hierarchy "theme" embodied by the *nexus* point. This point represents the fusion of a set of hierarchy-aware semantic sentences from specific to abstract that are relevant to the CoI $c$. Next, we detail the construction process.

**Querying the semantic hierarchy.** To obtain related super-/sub-categories, a semantic hierarchy $\mathcal{H}$ is crucial for our approach. In this study, we investigate two types of hierarchies: *i)* dataset-specific class taxonomies [6, 7, 10], and *ii)* hierarchies synthesized for the target test vocabulary using large language models (LLM). To generate the synthetic hierarchy, we follow Novack *et al.* [38] and query an LLM such as ChatGPT [39] to generate super-categories ($p = 3$) and sub-categories ($q = 10$) for each CoI $c \in \mathcal{C}^{\text{test}}$, creating a three-level hierarchy $\mathcal{H}$ (see App. B for details). With the hierarchy available, as depicted in Fig. 2(1), for each target CoI $c$, we retrieve *all* the related super-/sub-categories, which can assist in distinguishing $c$ from other concepts in the vocabulary across granularities [14]. Note that we exclude the root node (*e.g.*, "entity") from this process, as it does not help differentiate $c$ from other categories.

**Hierarchy-aware semantic sentence integration.** The collected categories contain both abstract and specific semantics useful for guiding the classification process. However, methods like simple ensembling [38] or concatenation [14] overlook some valuable knowledge *implicitly* provided by the hierarchy, namely the inherent internal relationships among concepts. Inspired by the hierarchy structure definition [53], we propose an **Is-A** connector to *explicitly* model these **interrelationships**. Specifically, for each target CoI $c$, the **Is-A** connector integrates the retrieved categories into sentences from the lowest sub-category (more specific) to the highest super-category (more abstract), including the target CoI name. As depicted in Fig. 2(2), this process yields a set of $K$ hierarchy-aware sentences $\{e_k^c\}_{k=1}^K$. Each sentence $e_k^c$ contains knowledge that spans from specific to abstract, all related to the target CoI and capturing their inherent relationships, as

```
A wooden baseball bat, which is a baseball
bat, which is a bat, which is a sports
equipment.
```

where the sub-categories, target category, and super-categories are color-coded in red, green, and blue.

**Semantic hierarchy *Nexus* construction.** A *nexus* $\mathbf{n}_c \in \mathbb{R}^D$ serves as a unifying embedding that fuses the hierarchy-aware knowledge contained in the integrated sentences $\{e_k^c\}_{k=1}^K$. As shown in Fig. 2(3), we employ a frozen VLM [42] text encoder $\mathcal{E}_{\text{txt}}$ to translate the integrated sentences into the region-language aligned semantic space compatible with the downstream OvOD detector. The semantic hierarchy *nexus* for the CoI $c$ is then constructed by

aggregating these individual sentence embeddings as:

$$\mathbf{n}_c = \text{Aggregator}\left(\{\mathcal{E}_{\text{txt}}(e_k^c)\}_{k=1}^{K}\right) \quad (1)$$

where, by default, we employ a straightforward but effective **mean-aggregator** to compute the mean vector of the set of sentence embeddings. The goal of the aggregation process is to fuse the expressive and granularity-robust knowledge into the *nexus* vector, as a "theme", from the encoded hierarchy-aware sentences. Inspired by text classification techniques in Natural Language Processing (NLP) [15, 28, 50], we also introduce an alternative aggregator, where we perform SVD decomposition of the sentence embeddings and replace the mean vector with the principal eigenvector as $\mathbf{n}_c$. We study its effectiveness in Sec. 4.1 and provide a detailed description in App. C.4.

### 3.2. Zero-shot Transfer with SHiNe

As shown in Fig. 2(bottom), once the *nexus* points are constructed for each CoI in the target vocabulary, the SHiNe classifier $\mathbf{N}$ can be directly applied to the OvOD detector for inference, assigning class names to proposed regions as:

$$\hat{c}_m = \arg\max_{c \in \mathcal{C}^{\text{test}}} \langle \mathbf{n}_c, \mathbf{z}_m \rangle \quad (2)$$

where $\mathbf{z}_m$ is the $m$-th region embedding. Given that $\mathbf{n}_c \in \mathbb{R}^D$, it becomes evident from Eq. 2 that SHiNehas the same computational complexity as the vanilla name-based OvOD classifier. Let us note that SHiNe is not limited to detection, it can be adapted to open-vocabulary classification by substituting the region embedding $\mathbf{z}_m$ with an image one. We validate this claim by also benchmarking on ImageNet-1k [7]. We provide the pseudo-code and time complexity analysis of SHiNe in App. C.2 and App. C.3, respectively.

## 4. Experiments

Table 1. Evaluation dataset descriptions of iNatLoc and FSOD. Label granularity ranges from finest (F) to coarsest (C) levels.

| Gran | iNatLoc Level | iNatLoc # Classes | iNatLoc Label Example | FSOD Level | FSOD # Classes | FSOD Label Example |
|---|---|---|---|---|---|---|
| F | 6 | 500 | Cyprinus carpio | 3 | 200 | Watermelon |
| | 5 | 317 | Cyprinus | | | |
| | 4 | 184 | Cyprinidae | 2 | 46 | Fruit |
| | 3 | 64 | Cypriniformes | | | |
| | 2 | 18 | Actinopterygii | 1 | 15 | Food |
| C | 1 | 5 | Chordata | | | |

**Evaluation protocol and datasets.** We primarily follow the cross-dataset transfer evaluation (CDTE) protocol [73] in our experiments. In this scenario, the OvOD detector is trained on one dataset and then tested on other datasets in a zero-shot manner. This enables a thorough evaluation of model performance across diverse levels of vocabulary granularity. We conduct experiments on two detection datasets: iNaturalist Localization 500 (iNatLoc) [6] and Few-shot Object Detection dataset (FSOD) [10], which have ground-truth hierarchies for evaluating object labeling at multiple levels of granularity. iNatLoc is a fine-grained detection dataset featuring a consistent six-level label hierarchy based on the biological tree of life, along with bounding box annotations for its validation set. FSOD is assembled from OpenImages [27] and ImageNet [7], structured with a two-level label hierarchy. For a more comprehensive evaluation, we use FSOD's test split and manually construct one more hierarchy level atop its existing top level, resulting in a three-level label granularity for evaluation. Tab. 1 outlines the number of label hierarchy levels and the corresponding category counts for both datasets, accompanied by examples to demonstrate the semantic granularity. Detailed dataset statistics and their hierarchies are available in App. A. We use the mean Average Precision (mAP) at an Intersection-over-Union (IoU) threshold of 0.5 (mAP50) as our main evaluation metric. Additional experiments on COCO [30] and LVIS [19] under the open-vocabulary protocol [18] are provided in App. I.

Table 2. Training signal combinations. LVIS [19] and COCO [30] are used as strong box-level supervision. ImageNet-21k [7] (IN-21k) and the 997-class subset (IN-L) of ImageNet-21k that overlaps with LVIS are used as weak image-level supervision.

| Notation | Strong Supervision | Weak Supervision |
|---|---|---|
| **I** | LVIS | N/A |
| **II** | LVIS | IN-L |
| **III** | LVIS | IN-21k |
| **IV** | LVIS & COCO | IN-21k |

**Baseline detector.** In our experiments, we use the pre-trained Detic [72] method as the baseline detector, given its open-source code and strong performance. Detic is a two-stage OvOD detector that relies on CenterNet2 [71] and incorporates a frozen text classifier generated from the CLIP ViT-B/32 text encoder [42] using a prompt of the form: `"a {Class}"`. Detic uses both detection and classification data (image-class weak supervisory signals) for training. In our experiments, we explore and compare with Detic under four variants of supervisory signal combinations as shown in Tab. 2. We study a ResNet-50 [21] and a Swin-B [32] backbone pre-trained on ImageNet-21k-P [44].

**SHiNe implementation details.** To directly apply our method to the baseline OvOD detector, we use the CLIP ViT-B/32 [42] text encoder to construct the SHiNe classifier and directly apply it to the baseline OvOD detector, following the pipeline described in Sec. 3.1. We use the mean-aggregator by default. In our experiments, we employ and study two sources for the hierarchy: the ground-truth hierarchy structure provided by the dataset and a synthetic hierarchy generated by an LLM. We use the gpt-3.5-turbo model [39] as our LLM via its public API to produce a simple 3-level hierarchy (comprising one child and one parent

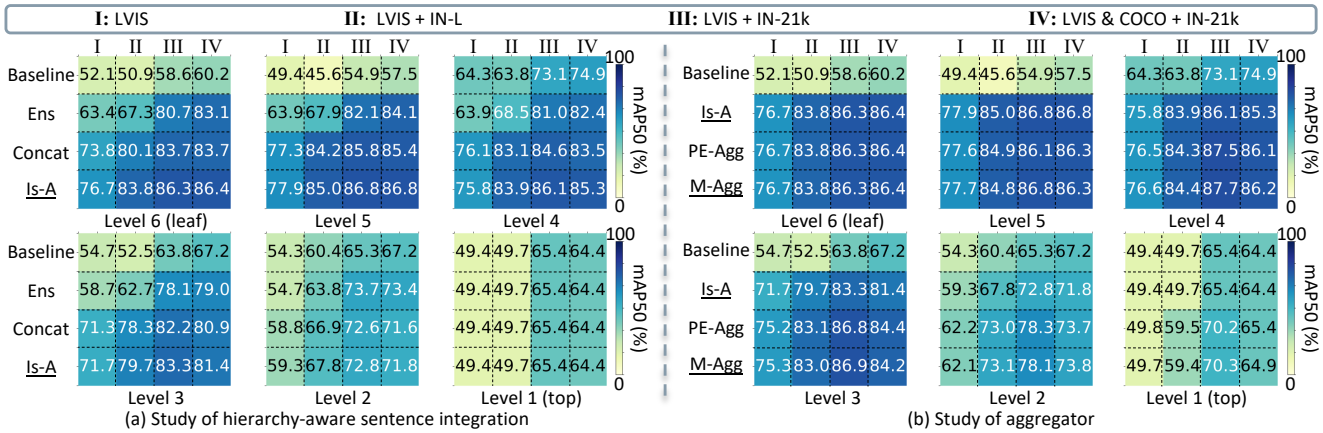


Figure 3. Study of hierarchy-aware sentence integration methods **(left)** and aggregators **(right)** across various label granularity levels on the iNatLoc dataset. Detic with a Swin-B backbone is used as the baseline. Darker background color indicates higher mAP50. The default components of SHiNe are underlined. Note that the experiment in (a) omits sub-categories and the aggregation step.

level) for the given target CoI vocabulary with temperature 0.7, as outlined in Sec. 3. We detail the hierarchy generation process in App. B and report the statistics.

### 4.1. Analysis of SHiNe

We first study the core components of SHiNe on the iNatLoc [6] using its *ground-truth* hierarchy. Consistent findings on the FSOD [10] dataset are reported in App. D.

**The Is-A connector effectively integrates hierarchy knowledge in natural sentences.** To assess the effectiveness of our **Is-A** connector, we design control experiments for constructing the OvOD classifier with a *single* sentence, omitting sub-categories and the aggregation step. Specifically, for a target CoI like `"Baseball bat"`, we retrieve only its super-categories at each ascending hierarchy level. We then explore three ways to integrate the CoI with its ascending super-categories in natural language and create the classifier vector as follows:

- **Ensemble (Ens)**: {`"baseball bat"`, `"bat"`, `"sports equipment"`}
- **Concatenate (Concat)**: `"A baseball bat bat sports equipment"`
- **Is-A (Ours)**: `"A baseball bat, which is a bat, which is a sports equipment"`

where the super-categories are colored in blue. For **Concat** and **Is-A**, we create the classifier vector for the target CoI by encoding the *single* sentence with the CLIP text encoder. For the **Ens** method, we use the average embedding of the ensembled names as: $\frac{1}{3}(\mathcal{E}_{\text{txt}}(\texttt{"baseball bat"}) + \mathcal{E}_{\text{txt}}(\texttt{"bat"}) + \mathcal{E}_{\text{txt}}(\texttt{"sports equipment"}))$. Next, we conduct control experiments to evaluate the three integration methods as well as the standard CoI name-based baseline methods. As shown in Fig. 3(a), except for the top levels where all methods degrade to the standard baseline (no super-category nodes), all methods outperform the baseline across all granularity levels by directing the model's focus towards more abstract concepts via the included super-categories. Among the methods compared, our **Is-A** connector excels across all granularity levels, boosting the baseline mAP50 by up to **+39.4** points (see last row and second column in Fig. 3(a-L5)). This underscores the effectiveness of our **Is-A** connector, which integrates related semantic concepts into sentences and explicitly models their relationships, yielding hierarchy-aware embeddings.

**A simple mean-aggregator is sufficient for semantic branch fusion.** We explored two aggregation methods: mean-aggregator (**M-Agg**) and principal eigenvector aggregator (**PE-Agg**). Note that in this experiment, all methods use the proposed **Is-A** connector to create *a set of* hierarchy-aware sentences to aggregate, ranging from each retrieved sub-category to the super-categories, as elaborated in Sec. 3. As Fig. 3(b) shows, both methods improve performance over the baseline across various models and label granularities. Note that these aggregators revert to the simple **Is-A** method at the leaf level where no sub-categories are available for aggregation. The benefits of aggregation methods are more pronounced with coarser granularity, significantly outperforming the baseline and the **Is-A** method, with gains up to **+9.8** on iNatLoc (see third row and second column in Fig. 3(b-L1). Notably, **M-Agg** generally outperforms **PE-Agg** despite its simplicity, making it the default choice for SHiNe in the subsequent experiments. Nonetheless, we aim to highlight the effectiveness of **PE-Agg**: to the best of our knowledge, this is the first study using the principal eigenvector as a classifier vector in vision-language models.

### 4.2. SHiNe on Open-vocabulary Detection

**SHiNe operates with different hierarchies.** In this section, we broaden our investigation to assess the effectiveness and the robustness of SHiNe with different semantic hierarchy sources. Tab. 3 shows the comparative analysis across various levels of label granularity between the

Table 3. Detection performance across varying label granularity levels, ranging from finest (F) to coarsest (C), on iNatLoc **(upper)** and FSOD **(lower)** datasets. SHiNe is directly applied to the baseline detector (BL) [72] with ground-truth (GT-H) and LLM-generated (LLM-H) hierarchies. ResNet-50 [21] **(left)** and Swin-B [32] **(right)** backbones [32] are compared. Four types of supervisory signal combinations are investigated. Note (†): At the L1-/L6-level of GT-H, no super-/sub-categories categories are used, respectively. mAP50 (%) is reported.

| Set | Gran | Level | ResNet-50 Backbone | | | | | | Swin-B Backbone | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | | **I** - LVIS | | | **II** - LVIS + IN-L | | | **III** - LVIS + IN-21k | | | **IV** - LVIS & COCO + IN-21k | | |
| | | | BL | SHiNe (GT-H) | SHiNe (LLM-H) | BL | SHiNe (GT-H) | SHiNe (LLM-H) | BL | SHiNe (GT-H) | SHiNe (LLM-H) | BL | SHiNe (GT-H) | SHiNe (LLM-H) |
| iNatLoc | (C ← F) | L6† | 32.0 | 48.4(+16.4) | **52.8(+20.8)** | 35.2 | 57.1(+21.9) | **58.3(+23.1)** | 58.6 | **86.3(+27.7)** | 84.5(+25.9) | 60.2 | **86.4(+26.2)** | 82.7(+22.5) |
| | | L5 | 28.2 | **49.4(+21.2)** | 41.1(+12.9) | 30.3 | **59.0(+28.7)** | 46.6(+16.3) | 54.9 | **86.8(+31.9)** | 76.3(+21.4) | 57.5 | **86.3(+28.8)** | 76.1(+18.6) |
| | | L4 | 40.1 | **51.5(+11.4)** | 50.4(+10.3) | 43.4 | **61.4(+18.0)** | 57.5(+14.1) | 73.1 | **87.7(+14.6)** | 84.0(+10.9) | 74.9 | **86.2(+11.3)** | 83.4(+8.5) |
| | | L3 | 38.8 | 56.5(+17.7) | **57.2(+18.4)** | 41.6 | **65.3(+23.7)** | 61.7(+20.1) | 63.8 | **86.9(+23.1)** | 83.6(+19.8) | 67.2 | **84.3(+17.1)** | 81.7(+14.5) |
| | | L2 | 34.4 | **45.0(+10.6)** | 43.9(+9.5) | 39.3 | **53.7(+14.4)** | 50.5(+11.2) | 65.3 | **78.1(+12.8)** | 77.2(+11.9) | 67.2 | 73.8(+6.6) | **74.5(+7.3)** |
| | | L1† | 31.6 | **33.6(+2.0)** | 33.5(+1.9) | 32.5 | **43.3(+10.8)** | 36.9(+4.4) | 65.4 | **70.3(+4.9)** | 63.8(-1.6) | 64.4 | **64.9(+0.5)** | 62.1(-2.3) |
| FSOD | (C ← F) | L3† | 49.7 | 52.1(+2.4) | **52.2(+2.5)** | 51.9 | 53.6(+1.7) | **53.7(+1.8)** | 66.0 | **66.7(+0.7)** | 66.3(+0.3) | 65.6 | **66.4(+0.8)** | **66.4(+0.8)** |
| | | L2 | 28.2 | **39.9(+11.7)** | 30.9(+2.7) | 27.8 | **39.8(+12.0)** | 29.8(+2.0) | 38.4 | **51.4(+13.0)** | 40.3(+1.9) | 39.4 | **52.4(+13.0)** | 41.5(+2.1) |
| | | L1† | 16.0 | **34.3(+18.3)** | 22.0(+6.0) | 16.5 | **31.4(+14.9)** | 21.0(+4.5) | 24.7 | **42.2(+17.5)** | 30.2(+5.5) | 25.0 | **42.5(+17.5)** | 29.6(+4.6) |

Table 4. Comparison with CoDet [33] and VLDet (VLD) [29] on iNatLoc and FSOD. SHiNe is applied to the baseline methods, respectively. All methods employ Swin-B [32] as backbone. Box-annotated LVIS [19] and image-caption-annotated CC3M [49] are used as supervisory signals. mAP50 (%) is reported.

| Set | Level | CoDet | SHiNe (GT-H) | SHiNe (LLM-H) | VLD | SHiNe (GT-H) | SHiNe (LLM-H) |
|---|---|---|---|---|---|---|---|
| iNatLoc | L6 | 48.7 | **80.1(+31.4)** | 75.1(+26.4) | 81.7 | **84.0(+2.3)** | 83.8(+2.1) |
| | L5 | 43.2 | **80.9(+37.7)** | 63.1(+19.9) | 83.7 | **84.7(+1.0)** | 82.1(-1.6) |
| | L4 | 64.0 | **80.5(+16.5)** | 73.8(+9.8) | 82.1 | 84.5(+2.4) | **85.8(+3.7)** |
| | L3 | 56.1 | **79.3(+23.2)** | 76.7(+20.6) | 77.7 | **83.9(+6.2)** | 83.3(+5.6) |
| | L2 | 61.3 | 65.3(+4.0) | **66.0(+4.7)** | 71.2 | 75.2(+4.0) | **77.2(+6.0)** |
| | L1 | 52.3 | **54.9(+2.6)** | 50.4(-1.9) | 66.1 | 66.7(+0.6) | **71.2(+5.1)** |
| FSOD | L3 | 60.5 | **62.5(+2.0)** | 61.6(+1.1) | 60.5 | **63.7(+3.2)** | 63.3(+2.8) |
| | L2 | 33.5 | **48.5(+15.0)** | 36.6(+3.1) | 33.9 | **49.2(+15.3)** | 37.4(+3.5) |
| | L1 | 19.9 | **39.7(+19.8)** | 25.4(+5.5) | 20.8 | **41.6(+20.8)** | 26.2(+5.4) |

baseline OvOD detector and our method, using either the ground-truth hierarchy or the LLM-generated hierarchy as proxies. We observe that our approach consistently surpasses the baseline by a large margin across all granularity levels on both datasets—and this holds true whether we employ the ground-truth or LLM-generated hierarchy. Averaged across all models and granularity levels on iNatLoc, our method yields an improvement of **+16.8** points using the ground-truth hierarchy and **+13.4** points with the LLM-generated hierarchy. For the FSOD dataset, we observe gains of **+10.3** and **+2.9** points, respectively. Although the performance gains are smaller with the LLM-generated hierarchy, they nonetheless signify a clear enhancement over the baseline across label granularities on all examined datasets. This shows that SHiNe is not reliant on ground-truth hierarchies. Even when applied to noisy, synthetic hierarchies, it yields substantial performance improvements. Additional results are in App. F and App. G.

**SHiNe operates with other OvOD detectors.** To evaluate SHiNe's generalizability, we apply SHiNe to additional OvOD detectors: CoDet [33] and VLDet (VLD) [29]. The evaluation results showcased in Tab. 4 affirm that SHiNe consistently improves the performance of CoDet and VLDet significantly across different granularities on both datasets, with both hierarchies. Further, we assess SHiNe on another DETR-style [5] detector, CORA [61], in App. H.

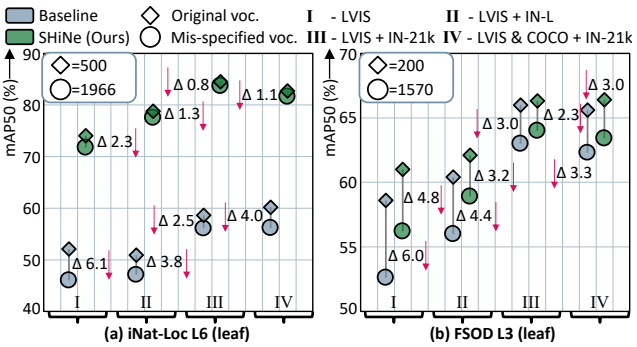


Figure 4. Analysis of OvOD detection performance under noisy *mis-specified* label vocabularies on iNatLoc **(left)** and FSOD **(right)** datasets. We assess the detection performance of both the baseline detector (in grey) and our method (in green) under varied supervision signals, contrasting results between the original (◇) and the expanded mis-specified (◯) vocabularies. SHiNe employs the LLM-generated hierarchy for both vocabularies. We report mAP50, highlighting the performance drop (Δ).

**SHiNe is resilient to *mis-specified* vocabularies.** In real-world applications, an authentic open vocabulary text classifier may be constructed using a vocabulary comprising a wide array of CoIs, even though only a subset of those specified classes appear in the test data. We define these as *mis-specified* vocabularies. Studying resilience in this challenging scenario is essential for practical applications. To this end, we gathered 500 class names from OpenImages [27]

Table 5. ImageNet-1k [7] zero-shot classification. We compare with two state-of-the-art hierarchy-based methods under WordNet (WDN) and LLM-generated hierarchies. Vanilla CLIP [42] serves as the baseline. We report top-1 accuracy, and FPS measured on the same NVIDIA RTX 2070 GPU with a batch size 1 and averaged over 10 runs. †: For fair comparison, we reproduce H-CLIP's results without its uncertainty estimation step and its *refined* WordNet hierarchy. In the original H-CLIP paper, a top-1 accuracy of 67.78% on ImageNet-1k was achieved using ViT-B/16 encoders.

| | | ViT-B/32 | | ViT-B/16 | | ViT-L/14 | |
|---|---|---|---|---|---|---|---|
| | | Acc(%) | FPS | Acc(%) | FPS | Acc(%) | FPS |
| | CLIP | 58.9 | 150 | 63.9 | 152 | 72.0 | 81 |
| WDN | H-CLIP† | 58.7(-0.2) | 3 | 63.8(-0.1) | 3 | 70.6(-1.4) | 2 |
| WDN | CHiLS | 59.6(+0.7) | 27 | 64.6(+0.7) | 28 | 72.1(+0.1) | 23 |
| WDN | SHiNe | **60.3(+1.4)** | 142 | **65.5(+1.6)** | 150 | **73.1(+1.1)** | 81 |
| LLM | H-CLIP | 55.8(-3.1) | 2 | 60.1(-3.8) | 2 | 66.9(-5.1) | 1 |
| LLM | CHiLS | 61.1(+2.2) | 26 | 66.1(+2.2) | 27 | 73.4(+1.4) | 23 |
| LLM | SHiNe | **61.6(+2.7)** | 141 | **66.7(+2.8)** | 149 | **73.6(+1.6)** | 81 |

and 1203 from LVIS [19], resulting in 1466 unique classes after deduplication. These are added as "noisy" CoIs to the iNatLoc and FSOD *leaf* label vocabularies, creating expanded sets with 1966 and 1570 CoIs, respectively. Using ChatGPT, SHiNe generates simple 3-level hierarchies for each class in these expanded vocabularies. As shown in Fig. 4, mis-specified vocabularies cause a decrease in baseline detector performance, dropping an average of **-4.1** points on iNatLoc and **-4.2** points on FSOD. However, interestingly, SHiNe not only continues to offer performance gains over the baseline detector but also mitigates the performance drop to **-1.4** on iNatLoc and **-3.3** on FSOD, respectively. This suggests that SHiNe not only improves the robustness but also enhances the resilience of the baseline detector when confronted with a mis-specified vocabulary.

### 4.3. SHiNe on Open-vocabulary Classification

In this section, we adapt SHiNe to open-vocabulary classification, by simply substituting the region embedding in Eq. 2 with an image embedding from the CLIP image encoder [42]. We evaluate it on the zero-shot transfer classification task using the well-established ImageNet-1k benchmark [7]. We compare SHiNe with two state-of-the-art hierarchy-based methods: CHiLS [38] and H-CLIP [14], which are specifically designed for classification.

**ImageNet-1k Benchmark.** In Tab. 5, we compare methods on ImageNet in terms of accuracy and frames-per-second (FPS). We observe that our approach consistently outperforms related methods. Comparing to the baseline that only uses class names, SHiNe improves its performance by an average of **+1.2%** and **+2.4%** across different model sizes using WordNet and LLM-generated hierarchies, respectively. Note that both CHiLS and H-CLIP introduce significant computational overheads due to their *search-on-the-fly* mechanism, resulting in a considerable decrease in inference speed. Consequently, this limits their scalability to detection tasks that necessitate per-region proposal inference for each image. For example, when processing detection results for *one* image with 300 region proposals, the overhead caused by CHiLS and H-CLIP would increase by ≈300×. In contrast, SHiNe maintains the same inference speed as the baseline, preserving its scalability.

Table 6. BREEDS-structured [48] ImageNet-1k zero-shot classification (with varying granularity). All methods use the BREED hierarchy and use CLIP ViT-B/16. Top-1 accuracy (%) reported.

| Level | # Classes | CLIP | H-CLIP [14] | CHiLS [38] | SHiNe |
|---|---|---|---|---|---|
| L1 | 10 | 56.2 | 67.9 (+11.7) | **73.8 (+17.6)** | 50.4(-5.8) |
| L2 | 29 | 56.8 | **69.3 (+12.5)** | 67.2 (+10.4) | 60.9(+4.1) |
| L3 | 128 | 43.3 | **62.4 (+19.1)** | 62.2 (+18.9) | 54.7(+11.4) |
| L4 | 466 | 55.2 | 69.6 (+14.4) | 70.1 (+14.9) | **70.3(+15.1)** |
| L5 | 591 | 62.4 | 65.9 (+3.5) | 64.5 (+2.1) | **69.1(+6.7)** |
| L6 | 98 | 73.1 | 75.4 (+2.3) | 73.5 (+0.4) | **78.9(+5.8)** |

**BREEDS ImageNet Benchmark.** Next, we analyze different granularity levels within ImageNet as organized by BREEDS [48]. In Tab. 6, we observe that CHiLS and H-CLIP surpass SHiNe at coarser granularity levels (L1 to L3). This is largely attributed to the BREEDS-modified hierarchy, where specific sub-classes in the hierarchy precisely correspond to the objects present in the test data. Yet, our method yields more substantial performance improvements at finer granularity levels (L4 to L6). Overall, the performance gains exhibited by all three methods underscore the benefits of using hierarchy information for improving open-vocabulary performance across granularities.

## 5. Conclusion

Given the importance of the vocabulary in open-vocabulary object detection, the robustness to varying granularities becomes critical for off-the-shelf deployment of OvOD models. Our preliminary investigations uncovered notable performance variability in existing OvOD detectors across different vocabulary granularities. To address this, we introduced SHiNe, a novel method that utilizes semantic knowledge from hierarchies to build *nexus*-based classifiers. SHiNe is training-free and can be seamlessly integrated with any OvOD detector, maintaining linear complexity relative to the number of classes. We show that SHiNe yields consistent improvements over baseline detectors across granularities with ground truth and LLM-generated hierarchies. We also extend SHiNe to open-vocabulary classification and achieve notable gains on ImageNet-1k [7].

**Acknowledgements.** E.R. is supported by MUR PNRR project FAIR - Future AI Research (PE00000013), funded by NextGenerationEU and EU projects SPRING (No. 871245) and ELIAS (No. 01120237). M.L. is supported by the PRIN project LEGO-AI (Prot. 2020TA3K9N). We thank Diane Larlus and Yannis Kalantidis for their helpful suggestions. M.L. thanks Zhun Zhong and Margherita Potrich for their constant support.

## References


[1] Relja Arandjelović, Alex Andonian, Arthur Mensch, Olivier J. Hénaff, Jean-Baptiste Alayrac, and Andrew Zisserman. Three ways to Improve Feature Alignment for Open Vocabulary Detection. arXiv:2303.13518, 2023. 2

[2] Björn Barz and Joachim Denzler. Hierarchy-based Image Embeddings for Semantic Image Retrieval. In *WACV*, 2019. 3

[3] Luca Bertinetto, Romain Mueller, Konstantinos Tertikas, Sina Samangooei, and Nicholas A. Lord. Making Better Mistakes: Leveraging Class Hierarchies with Deep Networks. In *CVPR*, 2020. 3

[4] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D. Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language Models are Few-Shot Learners. In *NeurIPS*, 2020. 2, 3

[5] Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. End-to-end object detection with transformers. In *ECCV*, 2020. 7

[6] Elijah Cole, Kimberly Wilber, Grant Van Horn, Xuan Yang, Marco Fornoni, Pietro Perona, Serge Belongie, Andrew Howard, and Oisin Mac Aodha. On Label Granularity and Object Localization. In *ECCV*, 2022. 1, 2, 3, 4, 5, 6, 12, 13, 20

[7] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. ImageNet: a Large-Scale Hierarchical Image Database. In *CVPR*, 2009. 2, 3, 4, 5, 8, 12, 13, 16, 18

[8] Jia Deng, Alexander C. Berg, Kai Li, and Li Fei-Fei. What Does Classifying more than 10,000 Image Categories Tell Us? In *ECCV*, 2010. 3

[9] Yu Du, Fangyun Wei, Zihe Zhang, Miaojing Shi, Yue Gao, and Guoqi Li. Learning to Prompt for Open-Vocabulary Object Detection with Vision-Language Model. In *CVPR*, 2022. 2

[10] Qi Fan, Wei Zhuo, Chi-Keung Tang, and Yu-Wing Tai. Few-Shot Object Detection with Attention-RPN and Multi-Relation Detector. In *CVPR*, 2020. 2, 4, 5, 6, 12, 13, 19

[11] Christiane Fellbaum. *WordNet: an Electronic Lexical Database*. MIT Press, 1998. 3

[12] Chengjian Feng, Yujie Zhong, Zequn Jie, Xiangxiang Chu, Haibing Ren, Xiaolin Wei, Weidi Xie, and Lin Ma. PromptDet: Towards Open-vocabulary Detection using Uncurated Images. In *ECCV*, 2022. 2

[13] Andrea Frome, Greg S. Corrado, Jon Shlens, Samy Bengio, Jeff Dean, Marc'Aurelio Ranzato, and Tomas Mikolov. DeViSE: A Deep Visual-Semantic Embedding Model. In *NeurIPS*, 2013. 3

[14] Yunhao Ge, Jie Ren, Andrew Gallagher, Yuxiao Wang, Ming-Hsuan Yang, Hartwig Adam, Laurent Itti, Balaji Lakshminarayanan, and Jiaping Zhao. Improving Zero-shot Generalization and Robustness of Multi-modal Models. In *CVPR*, 2023. 1, 2, 3, 4, 8, 15

[15] Felipe L. Gewers, Gustavo R. Ferreira, Henrique F. de Arruda, Filipi N. Silva, Cesar H. Comin, Diego R. Amancio, and Luciano da F. Costa. Principal Component Analysis: A Natural Approach to Data Exploration. *ACM Computing Surveys (CSUR)*, 54(4):1–34, 2021. 5, 15

[16] Joshua Goodman. Classes for Fast Maximum Entropy Training. In *ICASSP*, 2001. 3

[17] Jindong Gu, Zhen Han, Shuo Chen, Ahmad Beirami, Bailan He, Gengyuan Zhang, Ruotong Liao, Yao Qin, Volker Tresp, and Philip Torr. A Systematic Survey of Prompt Engineering on Vision-Language Foundation Models. arXiv:2307.12980, 2023. 3

[18] Xiuye Gu, Tsung-Yi Lin, Weicheng Kuo, and Yin Cui. Open-vocabulary Object Detection via Vision and Language Knowledge Distillation. In *ICLR*, 2022. 1, 2, 5

[19] Agrim Gupta, Piotr Dollar, and Ross Girshick. LVIS: A Dataset for Large Vocabulary Instance Segmentation. In *CVPR*, 2019. 5, 7, 8, 13, 18

[20] Ibrahim Ethem Hamamci, Sezgin Er, Enis Simsar, Anjany Sekuboyina, Mustafa Gundogar, Bernd Stadlinger, Albert Mehl, and Bjoern Menze. Diffusion-Based Hierarchical Multi-Label Object Detection to Analyze Panoramic Dental X-rays. In *MICCAI*, 2023. 3

[21] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep Residual Learning for Image Recognition. In *CVPR*, 2016. 5, 7, 18

[22] 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. 1, 2, 3

[23] Prannay Kaul, Weidi Xie, and Andrew Zisserman. Multi-Modal Classifiers for Open-Vocabulary Object Detection. In *ICML*, 2023. 2

[24] Muhammad Uzair Khattak, Hanoona Rasheed, Muhammad Maaz, Salman Khan, and Fahad Shahbaz Khan. MaPLe: Multi-modal Prompt Learning. In *CVPR*, 2023. 3

[25] John C. Knight. Safety Critical Systems: Challenges and Directions. In *ICSE*, 2002. 2

[26] Weicheng Kuo, Yin Cui, Xiuye Gu, AJ Piergiovanni, and Anelia Angelova. F-VLM: Open-Vocabulary Object Detection upon Frozen Vision and Language Models. In *ICLR*, 2023. 1, 2

[27] Alina Kuznetsova, Hassan Rom, Neil Alldrin, Jasper Uijlings, Ivan Krasin, Jordi Pont-Tuset, Shahab Kamali, Stefan Popov, Matteo Malloci, Alexander Kolesnikov, Tom Duerig, and Vittorio Ferrari. The Open Images Dataset V4. *IJCV*, 128:1956–1981, 2020. 5, 7

[28] Yong H Li and Anil K. Jain. Classification of Text Documents. *The Computer Journal*, 41(8):537–546, 1998. 5, 15

[29] Chuang Lin, Peize Sun, Yi Jiang, Ping Luo, Lizhen Qu, Gholamreza Haffari, Zehuan Yuan, and Jianfei Cai. Interpreting Word Embeddings with Eigenvector Analysis. In *ICLR*, 2023. 1, 7, 18

[30] 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. 5, 13, 18

[31] Wei Lin, Leonid Karlinsky, Nina Shvetsova, Horst Possegger, Mateusz Kozinski, Rameswar Panda, Rogerio Feris, Hilde Kuehne, and Horst Bischof. MAtch, eXpand and

Improve: Unsupervised Finetuning for Zero-Shot Action Recognition with Language Knowledge. In *ICCV*, 2023. 3

[32] 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, 7, 16, 19, 20

[33] Chuofan Ma, Yi Jiang, Xin Wen, Zehuan Yuan, and Xiaojuan Qi. CoDet: Co-Occurrence Guided Region-Word Alignment for Open-Vocabulary Object Detection. In *NeurIPS*, 2023. 7, 18

[34] Margarita Martínez-Díaz and Francesc Soriguera. Autonomous Vehicles: Theoretical and Practical Challenges. *Transportation Research Procedia*, 33:275–282, 2018. 2

[35] Sachit Menon and Carl Vondrick. Visual Classification via Description from Large Language Models. In *ICLR*, 2023. 2, 3

[36] Matthias Minderer, Alexey Gritsenko, and Neil Houlsby. Scaling Open-Vocabulary Object Detection. In *NeurIPS*, 2023. 2

[37] Frederic Morin and Yoshua Bengio. Hierarchical Probabilistic Neural Network Language Model. In *International Workshop on Artificial Intelligence and Statistics*, 2005. 3

[38] Zachary Novack, Julian McAuley, Zachary Chase Lipton, and Saurabh Garg. CHiLS: Zero-Shot Image Classification with Hierarchical Label Sets. In *ICML*, 2023. 1, 2, 3, 4, 8, 12, 13, 14, 15

[39] OpenAI. ChatGPT: A Large-Scale GPT-3.5-Based Model. https://openai.com/blog/chatgpt, 2022. 2, 3, 4, 5, 12

[40] Shubham Parashar, Zhiqiu Lin, Yanan Li, and Shu Kong. Prompting Scientific Names for Zero-Shot Species Recognition. In *EMNLP*, 2023. 1, 3

[41] Sarah Pratt, Ian Covert, Rosanne Liu, and Ali Farhadi. What Does a Platypus Look Like? Generating Customized Prompts for Zero-shot Image Classification. In *ICCV*, 2023. 3

[42] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning Transferable Visual Models From Natural Language Supervision. In *ICML*, 2021. 1, 2, 3, 4, 5, 8, 15, 16

[43] Shuhuai Ren, Aston Zhang, Yi Zhu, Shuai Zhang, Shuai Zheng, Mu Li, Alex Smola, and Xu Sun. Prompt Pre-Training with Twenty-Thousand Classes for Open-Vocabulary Visual Recognition. In *NeurIPS*, 2023. 3

[44] Tal Ridnik, Emanuel Ben-Baruch, Asaf Noy, and Lihi Zelnik-Manor. ImageNet-21K Pretraining for the Masses. In *NeurIPS*, 2021. 5

[45] Karsten Roth, Jae Myung Kim, A. Sophia Koepke, Oriol Vinyals, Cordelia Schmid, and Zeynep Akata. Waffling around for Performance: Visual Classification with Random Words and Broad Concepts. In *ICCV*, 2023. 3

[46] Michael A. Ruggiero, Dennis P. Gordon, Thomas M. Orrell, Nicolas Bailly, Thierry Bourgoin, Richard C. Brusca, Thomas Cavalier-Smith, Michael D. Guiry, and Paul M. Kirk. A Higher Level Classification of All Living Organisms. *PLOS ONE*, 10(4):e0119248, 2015. 3

[47] Miguel E. Ruiz and Padmini Srinivasan. Hierarchical Text Categorization Using Neural Networks. *Information retrieval*, 5(1):87–118, 2002. 3

[48] Shibani Santurkar, Dimitris Tsipras, and Aleksander Madry. BREEDS: Benchmarks for Subpopulation Shift. In *ICLR*, 2021. 8

[49] 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. 7, 18

[50] Jamin Shin, Andrea Madotto, and Pascale Fung. Interpreting Word Embeddings with Eigenvector Analysis. In *NeurIPS, IRASL workshop*, 2018. 5, 15

[51] Su-Jin Shin, Seyeob Kim, Youngjung Kim, and Sungho Kim. Hierarchical Multi-Label Object Detection Framework for Remote Sensing Images. *Remote Sensing*, 12(17):2734, 2020. 3

[52] Manli Shu, Weili Nie, De-An Huang, Zhiding Yu, Tom Goldstein, Anima Anandkumar, and Chaowei Xiao. Test-Time Prompt Tuning for Zero-Shot Generalization in Vision-Language Models. In *NeurIPS*, 2022. 3

[53] Carlos N. Silla and Alex A. Freitas. A Survey of Hierarchical Classification across Different Application Domains. *Data Mining and Knowledge Discovery*, 22:31–72, 2011. 3, 4

[54] Chufeng Tan, Xing Xu, and Fumin Shen. A Survey of Zero Shot Detection: Methods and Applications. *Cognitive Robotics*, 1:159–167, 2021. 2

[55] Grant Van Horn, Oisin Mac Aodha, Yang Song, Yin Cui, Chen Sun, Alex Shepard, Hartwig Adam, Pietro Perona, and Serge Belongie. The iNaturalist Species Classification and Detection Dataset. In *CVPR*, 2018. 3

[56] Catherine Wah, Steve Branson, Peter Welinder, Pietro Perona, and Serge Belongie. The Caltech-UCSD Birds-200-2011 Dataset. Technical Report CNS-TR-2011-001, California Institute of Technology, 2011. 3

[57] Wenhao Wang, Yifan Sun, Wei Li, and Yi Yang. TransHP: Image Classification with Hierarchical Prompting. In *NeurIPS*, 2023. 3

[58] Feihong Wu, Jun Zhang, and Vasant Honavar. Learning Classifiers Using Hierarchically Structured Class Taxonomies. In *SARA*, 2005. 3

[59] Jianzong Wu, Xiangtai Li, Shilin Xu Haobo Yuan, Henghui Ding, Yibo Yang, Xia Li, Jiangning Zhang, Yunhai Tong, Xudong Jiang, Bernard Ghanem, and Dacheng Tao. Towards Open Vocabulary Learning: A Survey. *IEEE TPAMI*, 2024. 1, 2, 3

[60] 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

[61] 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, 7, 18

[62] An Yan, Yu Wang, Yiwu Zhong, Chengyu Dong, Zexue He, Yujie Lu, William Yang Wang, Jingbo Shang, and Julian McAuley. Learning Concise and Descriptive Attributes for Visual Recognition. In *ICCV*, 2023. 3

[63] 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 Pre-training for Open-world Detection. In *NeurIPS*, 2022. 2

[64] Yuhang Zang, Wei Li, Kaiyang Zhou, Chen Huang, and Chen Change Loy. Open-Vocabulary DETR with Conditional Matching. In *ECCV*, 2022.

[65] Alireza Zareian, Kevin Dela Rosa, Derek Hao Hu, and Shih-Fu Chang. Open-Vocabulary Object Detection Using Captions. In *CVPR*, 2021. 1, 2

[66] Gengyuan Zhang, Jisen Ren, Jindong Gu, and Volker Tresp. Multi-event Video-Text Retrieval. In *ICCV*, 2023. 3

[67] Hao Zhang, Feng Li, Xueyan Zou, Shilong Liu, Chunyuan Li, Jianwei Yang, and Lei Zhang. A Simple Framework for Open-Vocabulary Segmentation and Detection. In *ICCV*, 2023. 2

[68] Yiwu Zhong, Jianwei Yang, Pengchuan Zhang, Chunyuan Li, Noel Codella, Liunian Harold Li, Luowei Zhou, Xiyang Dai, Lu Yuan, Yin Li, and Jianfeng Gao. RegionCLIP: Region-based Language-Image Pretraining. In *CVPR*, 2022. 2

[69] Kaiyang Zhou, Jingkang Yang, Chen Change Loy, and Ziwei Liu. Conditional Prompt Learning for Vision-Language Models. In *CVPR*, 2022. 3

[70] Kaiyang Zhou, Jingkang Yang, Chen Change Loy, and Ziwei Liu. Learning to Prompt for Vision-Language Models. *IJCV*, 130(7):2337–2348, 2022. 3

[71] Xingyi Zhou, Vladlen Koltun, and Philipp Krähenbühl. Probabilistic two-stage Detection. arXiv:2103.07461, 2021. 5

[72] Xingyi Zhou, Rohit Girdhar, Armand Joulin, Philipp Krähenbühl, and Ishan Misra. Detecting Twenty-thousand Classes using Image-level Supervision. In *ECCV*, 2022. 1, 2, 3, 5, 7, 16, 17, 18, 19, 20

[73] Chaoyang Zhu and Long Chen. A Survey on Open-Vocabulary Detection and Segmentation: Past, Present, and Future. arXiv:2307.09220, 2023. 1, 2, 3, 5, 18