Open-vocabulary Attribute Detection

María A. Bravo Sudhanshu Mittal Simon Ging Thomas Brox
{bravoma,mittal,gings,brox}@cs.uni-freiburg.de
University of Freiburg, Germany
https://ovad-benchmark.github.io/

Abstract

Vision-language modeling has enabled open-vocabulary tasks where predictions can be queried using any text prompt in a zero-shot manner. Existing open-vocabulary tasks focus on object classes, whereas research on object attributes is limited due to the lack of a reliable attribute-focused evaluation benchmark. This paper introduces the Open-Vocabulary Attribute Detection (OVAD) task and the corresponding OVAD benchmark. The objective of the novel task and benchmark is to probe object-level attribute information learned by vision-language models. To this end, we created a clean and densely annotated test set covering 117 attribute classes on the 80 object classes of MS COCO. It includes positive and negative annotations, which enables open-vocabulary evaluation. Overall, the benchmark consists of 1.4 million annotations. For reference, we provide a first baseline method for open-vocabulary attribute detection. Moreover, we demonstrate the benchmark's value by studying the attribute detection performance of several foundation models.

1. Introduction

One of the main goals of computer vision is to develop models capable of localizing and recognizing an open set of visual concepts in an image. This has been the main direction for the recently proposed Open-Vocabulary Detection (OVD) task [50] for object detection, where the goal is to detect a flexible set of object classes that are only defined at test time via a text query. Classical supervised object detection methods are bound to predict objects from a fixed set of pre-defined classes, and extending them to a very large number of classes is limited by the annotation effort. OVD methods overcome this constraint by utilizing vision-language modeling to learn about novel objects using the weak supervision of image-text pairs.

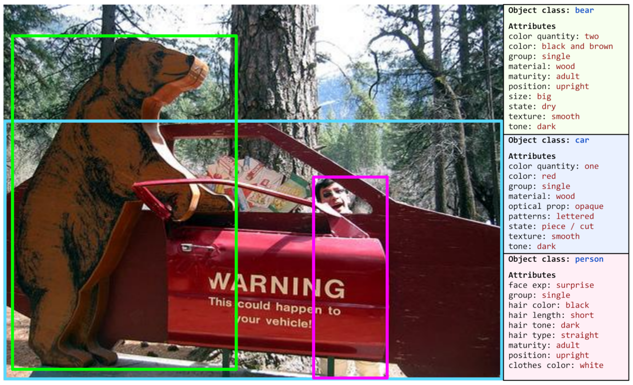

Figure 1. Example from the presented open vocabulary attribute detection benchmark. The objective is to detect all objects and visual attributes of each object in the image. Objects and attributes are only specified at test time via text prompts.

OVD methods for object detection have made fast progress and have even surpassed supervised baselines for rare (tail) classes [16]. Best OVD methods [16, 35, 53, 54] train with extra weak supervision using image classification datasets, which are focused on retrieving object information. However, it is unclear on how well OVD methods generalize information beyond the object class. This paper focuses on object-level attribute information, such as the object's state, size, and color.

Attributes play a significant role in an object's identity. A small change of an attribute in a description can modify our understanding of an object's appearance and perception. Imagine driving in a forest where you encounter a bear like the one in Figure 1. Even if you do not distinguish or know the type of bear, recognizing that it is made of wood is enough to realize that it is fake and harmless. A model capable of detecting object attributes enables a richer reasoning ability via combining objects and attributes. It allows the model to potentially extrapolate to novel object classes.

Acknowledgements This work was supported by the German Academic Exchange Service (DAAD) - 57440921 Research Grants - Doctoral Programmes in Germany, 2019/20. This work was partially funded by the German Research Foundation (DFG) - 401269959 and 417962828. We thank Philipp Schröppel, Silvio Galesso and Jan Bechtold for their critical feedback on the paper. We thank all OVAD annotators especially Mariana Sarmiento and Jorge Bravo for their help, time, and effort.

In this paper, we introduce the Open-Vocabulary Attribute Detection (OVAD) task. Its objective is to detect and recognize an open set of objects in an image together with an open set of attributes for every object. Both sets are defined by text queries during inference without knowledge of the tested classes during training. The OVAD task is a two-stage task. The first stage, referred to as open-vocabulary object detection [50], seeks to detect all objects in the image, including *novel* objects for which no bounding box or class annotation is available during training. The second stage seeks to determine all attributes present for each detected object. None of the attributes is annotated; therefore, all attributes are *novel*.

Testing the OVAD task requires an evaluation benchmark with unambiguous and dense attribute annotations to identify misses as well as false positive predictions. Current datasets [32, 33] for predicting attributes in-the-wild come with many missing or erroneous annotations, as discussed in more detail in Section 3.2. Thus, in this paper, we introduce the OVAD benchmark, an evaluation benchmark for open-vocabulary attribute detection. It is based on images of the MS COCO [29] dataset and only contains visually identifiable attributes. On average, the proposed benchmark has 98 attribute annotations per object instance, with 7.2 objects per image, for a total of 1.4 million attribute annotations, making it the most densely annotated object-level attribute dataset. It has a large coverage with 80 object categories and 117 attribute categories. It also provides negative attribute annotations, which enables quantifying false positive predictions. The benchmark is devoid of various labeling errors since it is manually annotated and quality-tested for annotation consistency. Our OVAD benchmark also extends the OVD benchmark [50] by including all 80 COCO object classes. This extension increases the novel set of objects from 17 to 32 classes. Together with the benchmark, we provide a first baseline method that learns the OVAD task to a reasonable degree. It learns the task from image-caption pairs by using all components of the caption, not only nouns. We also compare the performance of several off-the-shelf OVD models to get an insight of how much attribute information is implicitly comprised in nouns (e.g., *puppy* implies a young dog).

Moreover, we demonstrate the value of the benchmark by evaluating object-level attribute information learned by several open-source vision-language models, sometimes also referred to as foundation models, including CLIP [34], Open CLIP [20], BLIP [26], ALBEF [27], and X-VLM [51]. Such models learn from the weak supervision of image-text pairs, which is assumed to be available particularly via web content. The results show the extent to which the present success of foundation models on object classes generalizes to attributes.

Contributions (1) We introduce the **O**pen-**V**ocabulary **A**ttribute **D**etection (OVAD) task, where the objective is to detect all objects and predict their associated attributes. These objects and attributes belong to an open set of classes and can be queried using textual input. (2) We propose the OVAD benchmark: a clean and densely annotated evaluation dataset for open-vocabulary attribute detection, which can be used to evaluate open-vocabulary methods as well as foundation models. (3) We provide an attribute-focused baseline method for the OVAD task, which outperforms the existing open-vocabulary models that only aim for the object classes. (4) We test the performance of several open-source foundation models on visual attribute detection.

2. Related Work

Attribute prediction Several works have pursued the attribute prediction task to learn fine-grained information at different levels. Initial works focused on describing parts of the objects as attributes [13, 14]. In contrast to this partonomy identification, which can be regarded as a form of object detection (part detection), we focus on visual attributes represented by adjectives in human language. Other benchmarks for learning fine-grained semantics [21, 48] focus on tasks within narrow class domains, such as shoes [48], clothes [3,18], birds [45], and animals [46]. Another line of work [1,42,47] focuses on zero-shot object classification by inferring the attributes of an object as an intermediate step or relying on object-attribute compositionality [1, 8, 28] for zero-shot attribute-object classification. This work aims to evaluate the ability of vision-language models to detect and discriminate object-level attributes in a zero-shot manner.

Attribute detection benchmarks Recent works predict attributes in an open-domain setting, also known as "in-the-wild" setting. A few benchmarks have been proposed in this direction, along with some baseline methods. COCO Attributes [32] was the first such large-scale benchmark that annotated visual attributes for the COCO dataset. However, this dataset is limited in scope, with annotations only across 29 object categories. Visual Genome [25] offers a much wider coverage of attribute categories with more than 68 k attribute categories, including synonyms, but it contains very few attribute annotations for each object (0.74 attributes per instance). In Visual Genome, attribute annotations are not dense or exhaustive for every object since they were extracted from scene graph annotations which contain free-written form descriptions. Its sparsity, noise, and lack of negative annotations make it unsuitable for evaluating the OVAD task. Other works have introduced visual question answering datasets [2,25] with questions that require an understanding of vision, language, and common sense to respond. Even though the answers to these questions overlap with our objective (*e.g.* by asking about colors or materials), the performance on attributes and nouns cannot be isolated

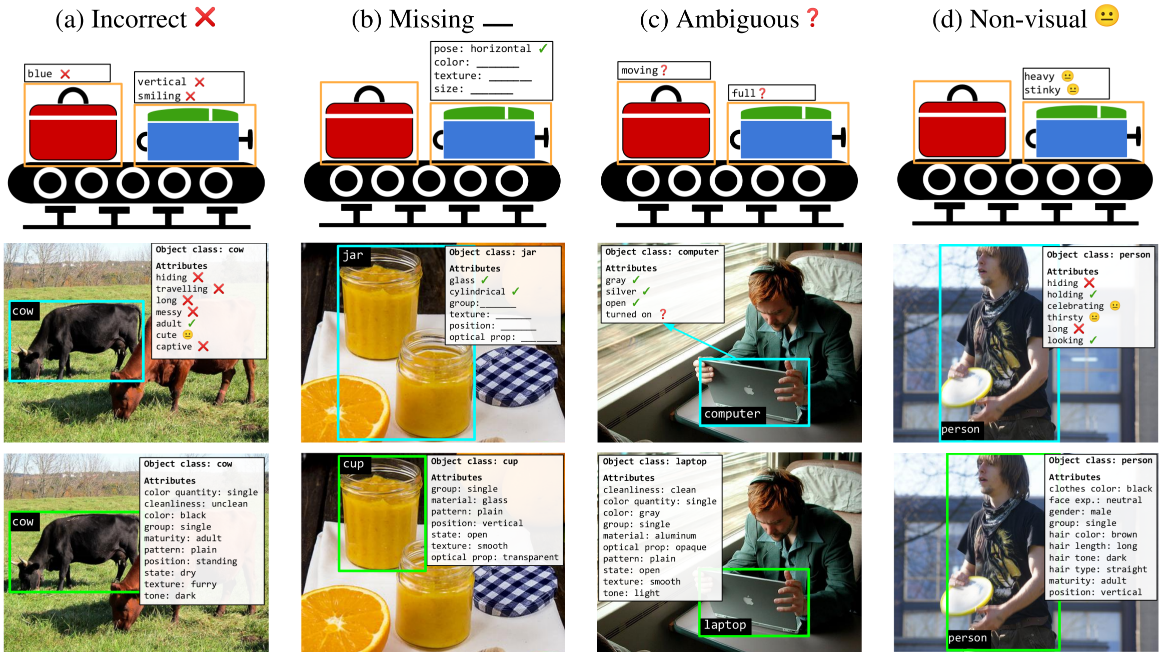

Figure 2. Four major types of errors prominent in previous attribute benchmarks with examples and their improved version in the proposed benchmark (last row). The top row shows a symbolic image with an example of how a briefcase and a trolley bag kept on a conveyor belt can be incorrectly marked with different types of errors. The second row of images shows examples from previous attribute benchmarks containing these errors. The last row shows examples from our proposed OVAD benchmark.

and analyzed using these datasets. VAW [33] proposed a large-scale dataset covering a wider range of attribute and object categories. They provide reliable positive and negative attribute labels for object instances and ensure that a minimum of 50 instances exist for each object-attribute pair. However, automated filtering techniques are used to keep the annotation cost feasible, resulting in very sparse annotations in terms of the number of instances per image and attributes per instance. Open Images [24] is a dataset consisting of 9 million images with image-level labels and bounding boxes. It provides attribute annotations for 288 object categories; however, it is limited to only 15 attribute categories that are not densely annotated for each object. We find that these benchmarks are of limited use for the precise evaluation and analysis of OVAD task. Therefore, in this work, we propose a new evaluation benchmark for attribute detection with clean and dense attribute annotations.

Open-vocabulary methods Zarenian *et al.* [50] introduced the open-vocabulary object detection problem, where the goal is to detect an open set of classes, some annotated (base classes) during training and others only defined at test time (novel classes). In this setting, the model learns in a weakly-supervised manner using image-caption pairs along with the annotations of base object classes. Various follow-up works [6, 15, 16, 35, 54] have improved the performance of open-vocabulary object detection. Bravo *et al.* [6] proposed a localized image-caption matching technique. Gu *et al.* [16] proposed an improved model using a pre-trained open-vocabulary classification model [34], created a new benchmark on LVIS [17], and showed some initial qualitative examples of fine-grained object detection. Recently, Zhou *et al.* [54] trained the classifier module of the detector by using extra class annotations. Other works [15, 35, 52] used pseudo-bounding-box annotations of base and novel classes to train their detector. In this work, we expand this problem formulation to include attributes.

Vision-language models have changed the way of approaching semantic learning tasks in computer vision by enabling the usage of large-scale free annotated data from the web. These foundation models [11, 20, 22, 23, 26, 27, 34, 51] use cross-modal objectives to learn to align visual concepts to their language representation leading them to achieve state-of-the-art performance on visual reasoning tasks. In this paper, we challenge five state-of-the-art vision-language models on the fine-grained task of open-vocabulary attribute detection.

3. Open-vocabulary Attribute Detection

3.1. The OVAD Task

Open-vocabulary attribute detection has a two-fold objective: (1) object detection and (2) discovery of attributes for all detected objects. Both object detection and attribute detection are formulated as open-vocabulary tasks. The first is known as open-vocabulary detection (OVD).

In previous work [50], OVD considers two disjoint sets of object classes - base $\mathcal{O}^B$ and novel $\mathcal{O}^N$ classes. The class labels and bounding boxes are given for the first set $\mathcal{O}^B$ during training, whereas the second set $\mathcal{O}^N$ needs to be derived automatically from image-caption pairs. Only at test time the set $\mathcal{O}^N$ is revealed. To be compatible with this setting from the literature, we use the object detection part of the OVAD task in the same way.

In contrast, for the second objective of OVAD task, none of the attributes are known during training. Rather all knowledge about attributes must be derived from image-caption pairs or pretrained vision-language models. Only at test time, the set of tested-visual attributes $\mathcal{A}$ is revealed. Using knowledge about the tested set of attribute classes for building the model violates the definition of the task.

Solving the task of OVAD requires the ability to detect both $\mathcal{O}^B$ and the unbounded $\mathcal{O}^N$ set of object classes as well as to determine whether attributes from $\mathcal{A}$ are present or absent for every object.

We also provide the OVAD task in a box-oracle setting, where the bounding box and object class annotations are available for all objects during inference. Thus, we only evaluate the second objective of the multi-label attribute detection task. This setting evaluates the attribute detection in isolation, independent of the mistakes made in the object detection part.

3.2. The OVAD Benchmark

For evaluating OVAD, it is necessary to have a benchmark dataset that contains annotations of both objects $\mathcal{O}$ and attributes $\mathcal{A}$. First, we discuss the limitations of previous datasets that provide both object and attribute annotations and then show how we rectify them for our benchmark.

Types of errors We identify four major sources of annotation errors in previous datasets, which make them unsuitable for the OVAD benchmark. The boundaries between these error types are blurry. Figure 2 shows an example for each error type followed by our corrected version to give an intuition for each of them. We summarize them as follows:

- **Type-A Incorrect:** Objects with incorrect attribute annotations. As shown in Figure 2(a), the *cow* is marked incorrectly with *hiding, travelling, long*, etc.
- **Type-B Missing:** Objects missing attribute annotations. As shown in Figure 2(b), the *jar* has missing attributes such as *group, texture*, and *position*.
- **Type-C Ambiguous:** Attributes that cannot be marked using the given image due to incomplete information. Figure 2(c) shows a *bag* on the conveyor belt marked as *moving*, and a *computer* marked as *turned on*, in the top and middle row respectively. These attributes only become valid when considering temporal information or a front view of the computer.
- **Type-D Non-visual:** Attributes that cannot be marked using visual information. These attributes are often subjective such as certain emotions or states of mind and occur due to poor selection of the attribute set. As shown in Figure 2(d), the *person* is annotated as *celebrating* and *thirsty*.

We aim to overcome the above-mentioned limitations of previous datasets by selecting a good set of attribute classes that can be accurately annotated for all object categories and is visually non-ambiguous for most samples. Our OVAD evaluation benchmark comprises 2000 images randomly selected from the MS-COCO [29] val2017 set. To ensure a densely annotated dataset with a large number of object annotations in an image, we started our annotation process with the COCO [29] object detection benchmark. We added bounding boxes for missing objects, revised inaccurate boxes, and removed incorrect object annotations. As a result, we obtained 14,300 object instances for the attribute annotation process. We manually labeled each object instance with 117 attributes following strict annotation guidelines to avoid above mentioned errors. The OVAD benchmark dataset is designed as a test set to evaluate models' fine-grained open-vocabulary detection capabilities. It is neither designed for classical supervised training nor as a validation set, as both contradict the open-vocabulary paradigm.

Selection of attributes We extracted adjectives from the captions of the COCO Captions dataset [9] using a parts-of-speech detector [4]. We selected the adjectives that occurred at least ten times and grouped them by synonyms using WordNet [5], Collins English Dictionary, and Oxford English Dictionary. We retained the synonyms and manually removed abstract, action-based, and non-visual attributes, such as *peaceful, walking, thirsty, etc.*, as shown in Figure 2(c&d). We considered the 80 MS-COCO object classes and removed attribute classes for which no positive object-attribute example existed. After this process, our final set consists of 117 unique attribute categories. We built a taxonomy and identified 19 attribute types or superclasses corresponding to *color, pattern, material, maturity, cooking state* and 14 others. A detailed diagram of the attribute taxonomy is included in the Supplementary A.1.

Annotation process The OVAD benchmark is fully annotated by humans, as compared to other works [25, 32, 33]. This ensures accurate ground-truth labels. The annotation was done using the open-source annotation platform "Computer Vision Annotation Tool" (CVAT) [39]. The OVAD benchmark has all attributes marked either as *positive*, *negative*, or *unknown*. We use the attribute taxonomy and the attribute types during the annotation process. Most of the attributes are mutually exclusive within their attribute type, *e.g*., *pose* can be either *vertical* or *horizontal* but not both simultaneously. For every object, annotators were directed to select one of the attributes for every attribute type as positive or unknown. Given the exclusiveness property, all non-selected attributes within the same attribute type were marked as negatives or unknown, respectively. This produced dense annotations for every object and ensured that the missing type errors were diminished (see Figure 2(b)). Attributes marked as unknown are excluded during evaluation. The unknown option either refers to an unknown attribute for an instance or an in-between case, where a discrete label can not be assigned clearly. This helps rectify ambiguous type errors like in Figure 2(c). We manually excluded infeasible object-attribute combinations during annotation, such as *smiling cup* or *open person*, to avoid incorrect type errors shown in Figure 2(a) and speed up the annotation process. We include a detailed description of the annotation process in the Supplementary B.

Statistics The OVAD benchmark is a medium-scale benchmark with a total of 1,401,484 attribute annotations over 2000 images. It considers 117 attribute categories that span across 80 object categories with a total of 14,300 object instances. There are 122,998 positive and 1,278,486 negative attribute annotations in total and 172,760 attribute instances are marked as unknown. Table 1 shows a summary of the dataset statistics together with other attribute datasets. Since OVAD is exclusively an evaluation benchmark, the number of images is not comparable to the other datasets. The OVAD evaluation benchmark is densely annotated with 7.2 box annotations per image, compared to 3.6 instances per image in VAW. Our benchmark offers, on average, 96.8 attribute annotations per box, with a total of 700.7 attribute annotations per image. This is much larger than any other object-level attribute benchmark. The benchmark provides both positive and negative attribute annotations grouped into 19 types of attributes.

Dataset	**OVAD** (ours)	**VAW** [33]	**COCO-A** [32]	**VG** [25]
Purpose	Test	Train+Test	Train+Test	Train+Test
# categories				
Objects	80	2,260	29	33,877
Attributes	117	620	196	68,111
Negative Labels	Yes	Yes	No	No
# instances				
Objects	14,300	260,895	188,426	3.8M
Attribute	1.4M	0.9M	3.4M	2.8M
Images	2,000	72,274	84,044	108,077
# instances per image				
Objects	7.2	3.6	2.2	35
Attributes	700.7 (+)61 (-)639	12.83 (+)5.4 (-)7.4	41.08	26
# instances per box				
Attributes	96.8 (+)8.3 (-)88.5	3.56 (+)1.51 (-)2.05	18.33	0.74

Table 1. Statistics of object-level attribute benchmarks. OVAD is densely annotated as compared to other datasets. (+) and (-) indicate positive and negative attribute labels respectively.

Evaluation metric As discussed in Section 3.1, the OVAD task can be evaluated under two settings: (1) open-vocabulary detection and (2) box-oracle setting. In the open-vocabulary detection setting, each ground-truth object instance is matched with at most one object prediction. To qualify as a positive match, the detection must have an Intersection over Union (IoU [12]) $\geq$ 0.5 independent of the ground-truth class. For every ground-truth object, the prediction with maximum IoU overlap is considered as the matching predicted object. We evaluate attribute performance by comparing the attribute scores and labels of matching ground-truth and predicted objects. Following Veit *et al*. [44], in the case that a ground-truth object has no matching prediction (IoU < 0.5 for all predictions), all attributes are marked as absent. We calculate the average precision (AP) [12] for every attribute category independently and then average across categories (mAP) [12]. Additionally, for completeness, we evaluate mAP at 0.5 IoU for open-vocabulary object detection on the 80 class object set; we call this set OVD-80. We use the *Generalized* evaluation that considers the probability across all object classes (base and novel). In the box-oracle setting, the attribute mAP metric is directly evaluated for ground-truth bounding boxes in an object-class-agnostic manner.

4. OVAD Baseline Method

In this section, we provide a baseline method for the OVAD task. The objective is to learn a vision model that detects objects and their corresponding attributes in an open-vocabulary manner. Our OVAD-Baseline comprises two models: a frozen language model G and an object detector F based on Faster-RCNN [36], where we replace the classification head with a linear layer that projects the visual features to the language space produced by G. Following other works [16,35,53,54], we use CLIP [34] as the language model. We define $g_w = G(w)$ as the embedding representation of a text composed of one or more words w, and $f_b = F(I_b)$ as the embedding representation of a box-region b of an image I.

Visual-text matching Throughout the paper, we use image-text pairs for learning the vision language alignment. These pairs can correspond to images and captions, box-regions and class labels, or in a more general setting, any box-region and text. We use the cosine similarity

$$s_{w,b} = \sigma(\frac{g_w \cdot f_b}{|g_w||f_b|} \cdot \tau) \tag{1}$$

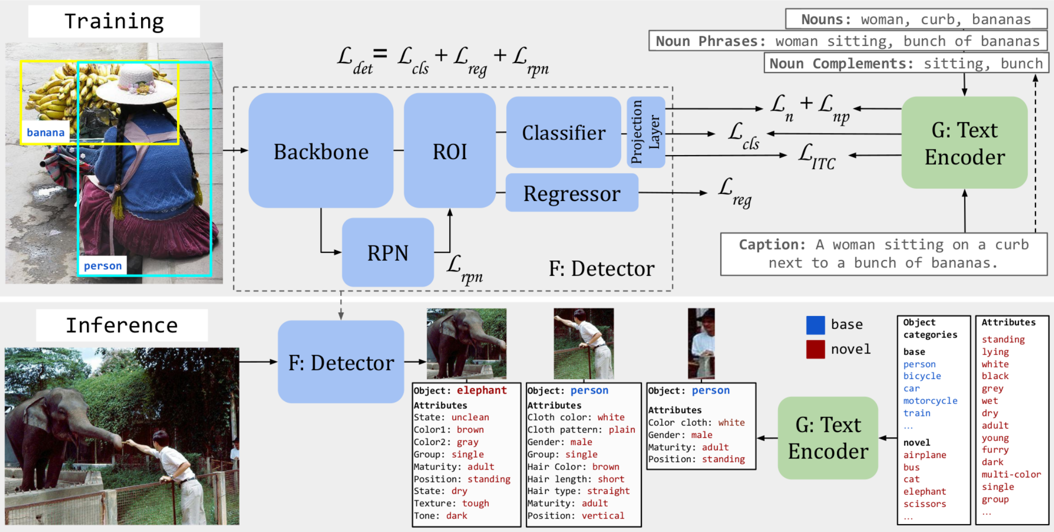

Figure 3. Training and inference setup for the OVAD-Baseline model. The method is a two-stage detector that matches image regions with text embeddings of nouns, noun phrases, noun complements, and captions. At inference, the detector detects the base and novel objects and their attributes by matching box-region embeddings with text embeddings of the object and attribute classes.

as matching score between a text w and a box-region b, where τ is a temperature hyper-parameter and σ corresponds to the sigmoid function.

Training objectives The detector F is trained with three objectives: 1) learn to localize objects in an image, 2) semantically match image representations with caption embeddings, and 3) train the classifier branch with proxy-labels to predict the novel classes and attributes.

For the first objective, we train F with labels and bounding box coordinates of the base classes O^B. We use the standard detection loss $\mathcal{L}_{det}$ from Faster R-CNN (shown in Figure 3) adapted for open-vocabulary. It comprises three losses: a region proposal network loss $\mathcal{L}_{rpn}$ [36], a class-agnostic l_1 loss as box regression loss $\mathcal{L}_{reg}$, and a similarity-based classification loss $\mathcal{L}_{cls}$ using the binary cross-entropy loss over the similarly score (1) between the visual embedding of the object box and the text embedding of the base classes.

For the second objective we use the image-text contrastive matching (ITC) loss

$$\mathcal{L}_{ITC} = -(y \log(s_{C,I}) + (1-y)\log(1-s_{C,I})), \quad (2)$$

with $s_{C,I}$ being the similarity score (1) between the image I and the caption C, and $y \in \{1, 0\}$ depending on whether I and C are a positive pair. We apply this loss to positive and negative image-caption pairs.

For the third objective, we match concepts within captions with image regions. These concepts, referred to as 'parts-of-caption' in this work, include nouns, noun phrases, and noun complements. They act as proxy-labels for objects and attributes. We obtain these parts-of-caption using a part-of-speech tagging method from the open-source software spaCy [19]. Nouns usually refer to object classes; however, they often reveal some attribute information, *e.g.*, *man/woman* are nouns that reveal gender, *cows* is a plural noun that reveals the quantity attribute. Noun phrases are usually adjective-noun combinations, which contain more explicit attribute information, such as *red helmet, wooden table*. We remove the nouns from the noun phrases to obtain "noun complements", which often contain adjectives, and use these to match directly with image regions. Since the location of these parts-of-caption is unknown, we match proxy-labels with the biggest predicted bounding box features $F(I_{b_{max}})$, similar to the usage of image labels in Detic's [54] training. Along with these positive pairs, we create negative proxy-labels using arbitrary image-caption pairs and apply the binary cross entropy loss (2). We refer to these losses as $\mathcal{L}_n$ and $\mathcal{L}_{np}$ for nouns and noun phrases/complements, respectively.

Inference During inference time, we consider a vocabulary composed of all object classes, $\mathcal{O}^B \cup \mathcal{O}^N$, together with the attribute classes $\mathcal{A}$ and use the language model G to get the corresponding text-vector representations of every class, as shown in Figure 3. We do not use any special text prompt for this purpose but consider all synonyms for every class (object/attribute) and average their text-vector representations. We obtain the final prediction for object and attribute classes by taking the sigmoid of the similarity (1) between the box-region representation $F(I_b)$ and the class-text embedding $G(c)$. We compute the output separately for each object and attribute class, predicting the class' presence or absence. See the supplementary for implementation details.

Method	**OVAD**				**Generalized OVD-80**		
	All	Head	Medium	Tail	Novel (32)	Base (48)	All (80)
Chance	8.6	36.0	7.3	0.6	-	-	-
OV-Faster-RCNN	11.7	34.4	13.1	1.9	0.3	53.3	32.1
VL-PLM [52]	13.2	32.6	16.3	2.6	19.7	58.8	43.2
Detic [54]	13.3	44.4	13.4	2.3	20.0	49.2	37.5
Rasheed *et al*. [35]	14.6	33.5	18.7	2.8	32.5	56.6	46.9
LocOv [6]	14.9	42.8	17.2	2.2	22.5	52.5	40.5
OVR [50]	15.1	46.3	16.7	2.1	17.9	51.8	38.2
OVAD-Baseline	**18.8**±0.3	**47.7**±0.6	**22.0**±0.5	**4.6**±0.5	24.7±0.6	49.1±0.2	39.3±0.4

Table 2. mAP for Open-vocabulary Attribute Detection (OVAD) and AP_{50} on Open-Vocabulary Detection (OVD-80).

box+cls $\mathcal{O}^B$	captions	nouns	noun phrases	noun comp.	OVAD mAP	AP_{50} - OVD-80 Novel (32)
✓					11.7±0.1	0.3±0.3
✓	✓				15.0±0.2	19.2±0.1
✓	✓	✓			16.2±0.3	23.2±0.8
✓	✓	✓	✓		15.9±0.1	23.7±0.5
✓	✓	✓		✓	18.8±0.3	24.7±0.6

Table 3. Text input ablation. OVAD and OVD-80 performance on novel classes using different types of text granularity as proxy-labels to train the model. *box+cls*: box and object-class labels for base objects, *noun phrases*: phrases that have one noun and some modifiers, *noun compl.*: noun phrases without the main noun. Training with finer granularity of text supervision is favorable.

5. Experiments

5.1. Open-vocabulary Attribute Detection

Open-vocabulary baseline methods We compare our OVAD-Baseline with previous off-the-shelf OVD models. For all methods base class object annotations come from MS COCO [29] 2017 training set and caption annotations from COCO Captions [9] 2017 training set. Given that OVD methods project the visual information to a language space, we use the similarity (1) of the visual representation of detected objects and the text embedding of every attribute to produce the attribute predictions, similar to the inference in Figure 3.

OV-Faster-RCNN is a Faster-RCNN adapted for open-vocabulary. Similar to OVAD-Baseline, the classification head of the detector is replaced with a linear layer to project the visual representation to the language space from the CLIP [34] text encoder. We train the detector network only using the class names of the base object classes and their box annotations. No caption was used for training.

OVR [50] and LocOv [6] train the object detector using two stages. First, the detectors learn a mapping between image regions and tokens in the caption via attention-based image-caption matching. OVR uses image grid-based regions for the matching, whereas LocOv introduces additional object proposal boxes. In the second stage, the models are fine-tuned using the base class annotations to learn the object detection task. Both models use BERT [10] as the text encoder.

Detic [54] and Rasheed *et al*. [35] train the detector using image-level labels filtered from the captions. Labels correspond to objects and are filtered using the class names of both base and novel classes, which technically is closed-vocabulary. Detic matches image-level labels, in text format, with the biggest box proposal. Rasheed *et al*. [35] first produce pseudo-labels for box proposals, using the image-level labels, to train the classification head of the detector. Similarly, VL-PLM [52] uses CLIP scores and from a class-agnostic object proposals to get pseudo-labels and train the OVD. All three models use CLIP [34] as the text encoder.

Results on the OVAD benchmark Table 2 presents results on the proposed OVAD benchmark for the six open-vocabulary detection methods. It shows results for attribute detection (OVAD) and object detection (OVD-80). Given that the attribute frequency has a long-tailed distribution and following previous works [17, 33], we report separate performances on attributes in the 'head', 'medium', and 'tail' of this distribution. These sets contain 16, 55, and 46 classes, respectively (see the supplementary for details).

All methods yield results above the chance level, even though the OVD methods were not designed to recognize attributes but only objects. Our OVAD-Baseline method outperforms these OVD methods. Methods that match image-regions with text-parts, either by using part-of-caption as in OVAD-Baseline or text tokens, as in OVR and LocOV, achieve better attribute mAP than those methods that use a single representation of the text for matching the image. Interestingly, methods that perform well on object detection are not necessarily better on the overall OVAD.

OVAD-Baseline ablation Table 3 breaks down the contributions of the parts-of-captions as proxy-labels to the performance of OVAD-Baseline. We find that using parts-of-caption as labels helps the model segregate the caption information, improving both the object and attribute detection performance. Training the model using noun complements makes the attribute supervision more explicit and makes the best use of the compositionality of the language structure.

5.2. Foundation Models Applied to Attributes

To demonstrate the value of an attribute evaluation benchmark, we tested the zero-shot performance of five pretrained vision-language models on attributes. To focus on attributes, we use the box-oracle setting. We crop the objects using their ground-truth bounding boxes and evaluate the attribute detection for each object instance independently. Our selection of models was based on the availability of code and model weights. Moreover, we selected models that process the text and the image independently, such that the matching score can be computed using the cosine similarity between the two representations.

All methods in Table 4 contain two transformer models that process image and text independently and use the

Method	**Training Data**	**OVAD-Box**			
		All	Head	Medium	Tail
Chance	-	8.6	36.0	7.3	0.6
CLIP RN50 [34]	400M (9)	15.8	42.5	17.5	4.2
CLIP VIT-B16 [34]	400M (9)	16.6	43.9	18.6	4.4
Open CLIP RN50 [20]	12M (7b)	11.8	41.0	11.7	1.4
Open CLIP ViT-B16 [20]	400M (8b)	16.0	45.4	17.4	3.8
Open CLIP ViT-B32 [20]	2B (8c)	17.0	44.3	18.4	5.5
ALBEF [27]	4M (1a,3,4,7a)	15.6	43.1	17.3	3.7
ALBEF [27]	14M (1a,3,4,7)	15.3	43.7	17.1	3.0
ALBEF [27]	14M (1a,3,4,7) + ft(2)	21.0	44.2	23.9	9.4
BLIP [26]	14M (1a,3,4,7)	17.0	46.6	18.3	5.0
BLIP [26]	129M (1a,3,4,7,8a)	18.2	44.4	20.7	5.7
BLIP [26]	129M (1a,3,4,7,8a) + ft(1a)	24.3	51.0	28.5	9.7
X-VLM [51]	4M (1^*,3^*,4,7a)	25.9	**50.3**	32.0	9.8
X-VLM [51]	16M (1^*,3^*,4,5^*,6^*,7) + ft(2)	26.2	48.7	31.2	12.1
X-VLM [51]	16M (1^*,4^*,4,5^*,6^*,7)	**28.1**	49.7	**34.2**	**12.9**
OVAD-Baseline-Box	0.11M (1a,$1b^{*\text{base}}$)	21.4±0.4	48.0±0.5	26.9±0.6	5.2±0.5

Table 4. Open-vocabulary Attribute Detection results (mAP) for foundation models in the box-oracle setup (OVAD-Box). * The model uses the localization information in the annotations of this dataset. + ft: final fine-tuning pass on the captions of this dataset. Table 5 details the training datasets.

(#) Dataset	**#Images**	**#Captions**	**#Objects**	**#Regions**
(1a) COCO Captions [9]	0.12M	0.57M	-	-
(1b) COCO Objects [29]	0.12M	-	0.86M	-
(2) RefCOCO+ [49]	0.019M	-	-	0.14M
(3) VG [25]	0.10M	-	2.5M	5.4M
(4) SBU Captions [31]	1M	1M	-	-
(5) OpenImages [24]	1.7M	0.67M	4.4M	3.3M
(6) Objects365 [40]	1.8M	-	29M	-
(7a) CC-3M [41]	2.95M	2.95M	-	-
(7b) CC-12M [7]	11.1M	11.1M	-	-
(8a) LAION [38]	115M	115M	-	-
(8b) LAION [38]	400M	400M	-	-
(8c) LAION [37]	2B	2B	-	-
(9) CLIP 400M [34]	400M	400M	-	-

Table 5. Training set legend and statistics

image-text contrastive learning (ITC) loss to learn from image-text pairs. ALBEF [27], BLIP [26], and X-VLM [51] additionally include a cross-attention model and use the image-text matching (ITM) loss. ALBEF and X-VLM use the masked language modeling (MLM) objective [10] to predict masked tokens from the caption in a bidirectional manner. BLIP uses the language modeling (LM) objective [30] to generate the caption conditioned on the image in an autoregressive manner. All three methods use a combination of clean and noisy data for training. ALBEF learns from noisy data by generating pseudo-targets via an online ensemble model [43]. BLIP instead filters noisy data and generates new captions to learn the multimodal matching. X-VLM uses localized region-text pairs to learn the vision-language alignment at multiple granularities.

Results and discussion Table 4 shows the results of foundation models on zero-shot attribute detection. Three interesting behaviors become evident.

a) Attribute detection is a challenge for foundation models. Compared to zero-shot image classification, where foundation models report very good accuracy [20,26,27,34,51], the absolute performance on attributes is surprisingly low. For reference, we trained a supervised attribute detector via cross-validation on our evaluation dataset, which achieved 48.16±0.52 mAP despite using a small training dataset; see Supplementary C. Based on the results, foundation models seem to be biased toward object classes and do not pick up fine-grained aspects such as attributes.

b) Not only the quantity but also the quality of training data is important. When scaling from 400M to 2B image-text pairs, OpenCLIP improves by 6.25% for *All* attribute performance. BLIP improves by 7.06% when scaling it from 14M to 129M, and quadrupling the data improves X-VLM by 8.46%. However, the models only reach a good performance once they are further trained on curated data using only ITC and ITM objectives. For instance, ALBEF and BLIP improve their *All* attribute performance this way by 37.25% and 33.52%, respectively.

c) Localized image region-text matching helps vision-language alignment. X-VLM and OVAD-Baseline-Box use a localized image region-text matching objective compared to the other methods. X-VLM clearly outperforms all other methods, but it reduces its performance by 6.76% when fine-tuning for image-caption retrieval. OVAD-Baseline-Box outperforms foundation models trained on more than 3000 times larger noisy datasets (CLIP and OpenCLIP), and more than 1000 times larger datasets which include the same clean subset (Table 5(1a)) (ALBEF and BLIP in their pretrained version). In Table 3 OVAD-Baseline shows an increase in performance when using parts-of-caption for explicit visual-text matching during training. We believe that the success of both methods comes from the localized alignment between visual and text context, which is partially lost when specializing for image-caption retrieval.

6. Conclusion

We studied the ability of vision-language models to recognize attributes. To this end, we proposed the novel open-vocabulary attribute detection (OVAD) task and introduced the OVAD benchmark, a clean and densely annotated object-level attribute dataset for evaluating OVAD task. We provided a baseline method that exploits fine-grained information contained in captions, which outperforms OVD models for the OVAD task. Finally, we tested the performance of publicly available foundation models on attribute recognition. We found that the performance of these models on attributes stays clearly behind their performance on objects revealing a direction for further research.

References

[1] Ziad Al-Halah, Makarand Tapaswi, and Rainer Stiefelhagen. Recovering the missing link: Predicting class-attribute associations for unsupervised zero-shot learning. In *CVPR*, 2016. 2

[2] Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C Lawrence Zitnick, and Devi Parikh. Vqa: Visual question answering. In *ICCV*, 2015. 2

[3] Tamara L Berg, Alexander C Berg, and Jonathan Shih. Automatic attribute discovery and characterization from noisy web data. In *ECCV*, 2010. 2

[4] Steven Bird, Ewan Klein, and Edward Loper. *Natural language processing with Python: analyzing text with the natural language toolkit*. " O'Reilly Media, Inc.", 2009. 4

[5] Francis Bond and Kyonghee Paik. A survey of wordnets and their licenses. In *International Global Wordnet Conference*, 2012. 4

[6] Maria A. Bravo, Sudhanshu Mittal, and Thomas Brox. Localized vision-language matching for open-vocabulary object detection. In *GCPR*, 2022. 3, 7

[7] Soravit Changpinyo, Piyush Sharma, Nan Ding, and Radu Soricut. Conceptual 12m: Pushing web-scale image-text pretraining to recognize long-tail visual concepts. In *CVPR*, 2021. 8

[8] Hui Chen, Zhixiong Nan, Jingjing Jiang, and Nanning Zheng. Learning to infer unseen attribute-object compositions. *arXiv preprint arXiv:2010.14343*, 2020. 2

[9] Xinlei Chen, Hao Fang, Tsung-Yi Lin, Ramakrishna Vedantam, Saurabh Gupta, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco captions: Data collection and evaluation server. *arXiv preprint arXiv:1504.00325*, 2015. 4, 7, 8

[10] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018. 7, 8

[11] Zi-Yi Dou, Yichong Xu, Zhe Gan, Jianfeng Wang, Shuohang Wang, Lijuan Wang, Chenguang Zhu, Pengchuan Zhang, Lu Yuan, Nanyun Peng, et al. An empirical study of training end-to-end vision-and-language transformers. In *CVPR*, 2022. 3

[12] Mark Everingham, Luc Van Gool, Christopher K. I. Williams, John M. Winn, and Andrew Zisserman. The pascal visual object classes (voc) challenge. *IJCV*, 2010. 5

[13] Ali Farhadi, Ian Endres, Derek Hoiem, and David Forsyth. Describing objects by their attributes. In *CVPR*, 2009. 2

[14] Vittorio Ferrari and Andrew Zisserman. Learning visual attributes. In *NeurIPS*, 2007. 2

[15] Mingfei Gao, Chen Xing, Juan Carlos Niebles, Junnan Li, Ran Xu, Wenhao Liu, and Caiming Xiong. Open vocabulary object detection with pseudo bounding-box labels. In *ECCV*, 2022. 3

[16] Xiuye Gu, Tsung-Yi Lin, Weicheng Kuo, and Yin Cui. Open-vocabulary object detection via vision and language knowledge distillation. In *ICLR*, 2022. 1, 3, 5

[17] Agrim Gupta, Piotr Dollar, and Ross Girshick. Lvis: A dataset for large vocabulary instance segmentation. In *CVPR*, 2019. 3, 7

[18] Xintong Han, Zuxuan Wu, Phoenix X Huang, Xiao Zhang, Menglong Zhu, Yuan Li, Yang Zhao, and Larry S Davis. Automatic spatially-aware fashion concept discovery. In *ICCV*, 2017. 2

[19] Matthew Honnibal, Ines Montani, Sofie Van Landeghem, and Adriane Boyd. spacy: Industrial-strength natural language processing in python, 2020. 6

[20] Gabriel Ilharco, Mitchell Wortsman, Ross Wightman, Cade Gordon, Nicholas Carlini, Rohan Taori, Achal Dave, Vaishaal Shankar, Hongseok Namkoong, John Miller, Hannaneh Hajishirzi, Ali Farhadi, and Ludwig Schmidt. Openclip, 2021. 2, 3, 8

[21] Phillip Isola, Joseph J Lim, and Edward H Adelson. Discovering states and transformations in image collections. In *CVPR*, 2015. 2

[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. 3

[23] Wonjae Kim, Bokyung Son, and Ildoo Kim. Vilt: Vision-and-language transformer without convolution or region supervision. In *ICML*, 2021. 3

[24] Ivan Krasin, Tom Duerig, Neil Alldrin, Andreas Veit, Sami Abu-El-Haija, Serge Belongie, David Cai, Zheyun Feng, Vittorio Ferrari, Victor Gomes, Abhinav Gupta, Dhyanesh Narayanan, Chen Sun, Gal Chechik, and Kevin Murphy. Openimages: A public dataset for large-scale multi-label and multi-class image classification. *Dataset available from https://github.com/openimages*, 2016. 3, 8

[25] Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, Li-Jia Li, David A Shamma, Michael Bernstein, and Li Fei-Fei. Visual genome: Connecting language and vision using crowdsourced dense image annotations. *IJCV*, 2017. 2, 4, 5, 8

[26] Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *ICML*, 2022. 2, 3, 8

[27] Junnan Li, Ramprasaath R. Selvaraju, Akhilesh Deepak Gotmare, Shafiq Joty, Caiming Xiong, and Steven Hoi. Align before fuse: Vision and language representation learning with momentum distillation. In *NeurIPS*, 2021. 2, 3, 8

[28] Yong-Lu Li, Yue Xu, Xiaohan Mao, and Cewu Lu. Symmetry and group in attribute-object compositions. In *CVPR*, 2020. 2

[29] 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. 2, 4, 7, 8

[30] Andriy Mnih and Geoffrey E Hinton. A scalable hierarchical distributed language model. *NeurIPS*, 2008. 8

[31] Vicente Ordonez, Girish Kulkarni, and Tamara L. Berg. Im2text: Describing images using 1 million captioned photographs. In *NeurIPS*, 2011. 8

[32] Genevieve Patterson and James Hays. Coco attributes: Attributes for people, animals, and objects. In *ECCV*, 2016. 2, 4, 5

[33] Khoi Pham, Kushal Kafle, Zhe Lin, Zhihong Ding, Scott Cohen, Quan Tran, and Abhinav Shrivastava. Learning to predict visual attributes in the wild. In *CVPR*, 2021. 2, 3, 4, 5, 7

[34] 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, 5, 7, 8

[35] Hanoona Abdul Rasheed, Muhammad Maaz, Muhammd Uzair Khattak, Salman Khan, and Fahad Khan. Bridging the gap between object and image-level representations for open-vocabulary detection. In *NeurIPS*, 2022. 1, 3, 5, 7

[36] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. In *NeurIPS*, 2015. 5, 6

[37] Christoph Schuhmann, Romain Beaumont, Richard Vencu, Cade W Gordon, Ross Wightman, Mehdi Cherti, Theo Coombes, Aarush Katta, Clayton Mullis, Mitchell Wortsman, Patrick Schramowski, Srivatsa R Kundurthy, Katherine Crowson, Ludwig Schmidt, Robert Kaczmarczyk, and Jenia Jitsev. LAION-5b: An open large-scale dataset for training next generation image-text models. In *NeurIPS Datasets and Benchmarks Track*, 2022. 8

[38] Christoph Schuhmann, Robert Kaczmarczyk, Aran Komatsuzaki, Aarush Katta, Richard Vencu, Romain Beaumont, Jenia Jitsev, Theo Coombes, and Clayton Mullis. Laion-400m: Open dataset of clip-filtered 400 million image-text pairs. In *NeurIPS Workshop Datacentric AI*, 2021. 8

[39] Boris Sekachev, Nikita Manovich, Maxim Zhiltsov, Andrey Zhavoronkov, Dmitry Kalinin, Ben Hoff, TOsmanov, Dmitry Kruchinin, Artyom Zankevich, DmitriySidnev, Maksim Markelov, Johannes222, Mathis Chenuet, a andre, telenachos, Aleksandr Melnikov, Jijoong Kim, Liron Ilouz, Nikita Glazov, Priya4607, Rush Tehrani, Seungwon Jeong, Vladimir Skubriev, Sebastian Yonekura, vugia truong, zliang7, lizhming, and Tritin Truong. opencv/cvat: v1.1.0, 2020. 4

[40] 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. 8

[41] 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. 8

[42] Tristan Sylvain, Linda Petrini, and Devon Hjelm. Locality and compositionality in zero-shot learning. In *ICLR*, 2020. 2

[43] Antti Tarvainen and Harri Valpola. Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results. *NeurIPS*, 2017. 8

[44] Andreas Veit, Neil Alldrin, Gal Chechik, Ivan Krasin, Abhinav Gupta, and Serge Belongie. Learning from noisy large-scale datasets with minimal supervision. In *CVPR*, 2017. 5

[45] Peter Welinder, Steve Branson, Takeshi Mita, Catherine Wah, Florian Schroff, Serge Belongie, and Pietro Perona. Caltech-ucsd birds 200. Technical report, 2010. 2

[46] Yongqin Xian, Bernt Schiele, and Zeynep Akata. Zero-shot learning-the good, the bad and the ugly. In *CVPR*, 2017. 2

[47] Wenjia Xu, Yongqin Xian, Jiuniu Wang, Bernt Schiele, and Zeynep Akata. Attribute prototype network for zero-shot learning. In *NeurIPS*, 2020. 2

[48] Aron Yu and Kristen Grauman. Fine-grained visual comparisons with local learning. In *CVPR*, 2014. 2

[49] Licheng Yu, Patrick Poirson, Shan Yang, Alexander C Berg, and Tamara L Berg. Modeling context in referring expressions. In *ECCV*, 2016. 8

[50] Alireza Zareian, Kevin Dela Rosa, Derek Hao Hu, and Shih-Fu Chang. Open-vocabulary object detection using captions. In *CVPR*, 2021. 1, 2, 3, 4, 7

[51] Yan Zeng, Xinsong Zhang, and Hang Li. Multi-grained vision language pre-training: Aligning texts with visual concepts. In *ICML*, 2022. 2, 3, 8

[52] Shiyu Zhao, Zhixing Zhang, Samuel Schulter, Long Zhao, BG Vijay Kumar, Anastasis Stathopoulos, Manmohan Chandraker, and Dimitris N Metaxas. Exploiting unlabeled data with vision and language models for object detection. In *ECCV*, 2022. 3, 7

[53] 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. 1, 5

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