

Open-Set Recognition in the Age of Vision-Language Models

Dimity Miller, Niko Sünderhauf, Alex Kenna, and Keita Mason

Queensland University of Technology, Brisbane, Australia**
{d24.miller, niko.suenderhauf}@qut.edu.au

Abstract. Are vision-language models (VLMs) for open-vocabulary perception inherently open-set models because they are trained on internet-scale datasets? We answer this question with a clear *no* – VLMs introduce closed-set assumptions via their finite query set, making them vulnerable to open-set conditions. We systematically evaluate VLMs for open-set recognition and find they frequently misclassify objects not contained in their query set, leading to alarmingly low precision when tuned for high recall and vice versa. We show that naively increasing the size of the query set to contain more and more classes does *not* mitigate this problem, but instead causes diminishing task performance *and* open-set performance. We establish a revised definition of the open-set problem for the age of VLMs, define a new benchmark and evaluation protocol to facilitate standardised evaluation and research in this important area, and evaluate promising baseline approaches based on predictive uncertainty and dedicated negative embeddings on a range of open-vocabulary VLM classifiers and object detectors.

1 Introduction

In 2013, Scheirer et al. [37] described the closed-set assumption ingrained in almost all vision models at the time: all test classes are known during training and the model is never exposed to novel or unknown classes during testing. Real-world applications challenge this assumption and test in open-set conditions – encountering unexpected objects not in the training dataset – and causing potentially dangerous overconfident misclassifications [4, 10, 26, 38]. Open-set recognition [37] was introduced to address this issue by evaluating the ability of vision models to identify and reject open-set inputs as unknown.

Vision-language models (VLMs) have recently revolutionised the fields of image classification [14, 19, 33] and object detection [15, 48, 50]. Trained on vast and diverse internet-scale datasets, VLMs recognise an extensive variety of classes and *seemingly* are open-set by default [42]. Our paper challenges this view.

In the age of VLMs, we introduce an updated definition of the open-set problem – while previous definitions focused on a finite *training* set [37], we show that

** The authors acknowledge ongoing support from the QUT Centre for Robotics.

Evaluation code is publicly available at github.com/dimitymiller/openset_vlms

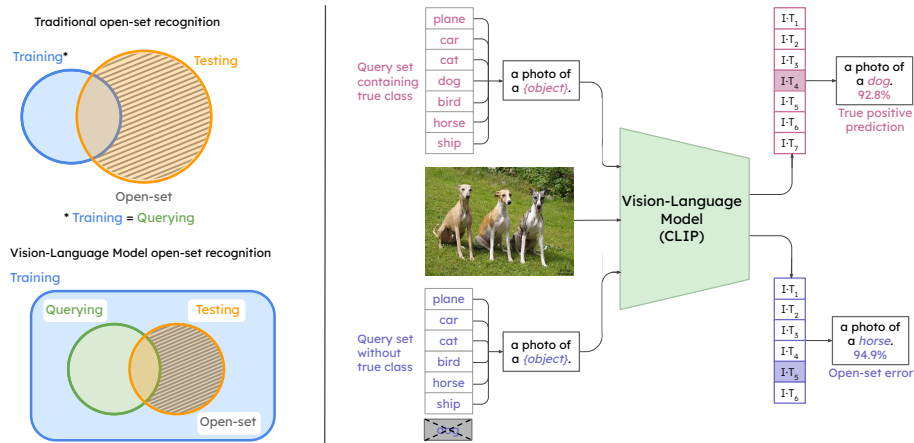


Fig. 1: Left: Traditional open-set recognition arises when testing classes outside a model’s finite training class set. While VLMs are trained on internet-scale datasets containing most conceivable object classes, they use a finite query set for classification. Open-set recognition arises when test classes are not included in the query set. Right: When testing an object present in the predefined query set, VLMs often correctly classify the object. When testing an object that is not present in the query set (i.e. an open-set object), VLMs often misclassify the object as a query class with high confidence (i.e. an open-set error).

VLMs impose closed-set assumptions through a finite *query* set. As illustrated in Fig. 1, VLMs for open-vocabulary image classification or object detection are still evaluated under closed-set conditions: they compare image embeddings with text embeddings from a predefined, finite query set of class labels. This introduces a closed-set assumption that all classes encountered during testing are included in this query set. Open-set conditions emerge when VLMs encounter objects that are not included in the query set (unknown objects). As we show in Section 5, even state-of-the-art VLMs heavily degrade in performance and misclassify unknown objects as belonging to the query set with high confidence.

In theory, this problem disappears if a VLM’s query set contains *every possible class label in a vocabulary*. However, we show that naively increasing the query set to contain more and more class labels results in worse task performance, where increasing numbers of misclassifications occur (see Sec. 5.4). The serious concerns about deploying vision models in safety-critical applications voiced in the pre-VLM era [2, 38] continues to be relevant and underscore the urgent need for further research into the open-set problem for VLMs.

Our paper is the first to systematically investigate and quantify the open-set vulnerabilities of VLM-based open-vocabulary object detectors and zero-shot image classifiers. We re-define the open-set problem for VLMs (Section 3) and compare promising baseline strategies, leveraging predictive uncertainty (Section 5.1) and different approaches to inserting dedicated **negative** (unknown) class representations into the VLM pipeline in Section 5.2. We propose a new

benchmark and evaluation protocol in Section 4 to foster further research and evaluation of VLMs for open-set recognition.

2 Background

2.1 Vision-Language Models for Zero-shot Classification

A recently emerged foundation model is the *vision-language model (VLM)*, which learns a multi-modal feature space that can jointly represent visual features and text [1, 19, 33, 46, 51] (and in some cases other modalities as well [14, 47]). Large-scale contrastive, self-supervised learning on internet-scale datasets provides VLMs with a rich feature representation that can be adapted for many downstream tasks, one of which is zero-shot or open vocabulary image classification [1, 14, 19, 33, 46, 47, 51]. Despite not being trained for classification tasks or datasets, VLMs generalise surprisingly well and achieve near state-of-the-art supervised learning performance [46]. Alongside this, a distinguishing advantage of VLMs is their adaptability to different datasets with different classes – the set of classes can be changed at will depending on the dataset [19, 33].

2.2 Vision-Language Models for Open Vocabulary Object Detection

Unlocked by the advances in VLMs for zero-shot classification, [50] proposed open vocabulary object detection, where detectors must generalise to an arbitrary set of object classes at test time, even object classes that were not seen during training [50]. This is related to our proposed open-set recognition, but assumes that novel object classes are added to the query set, whereas we test against object classes that are not present in the query set.

Existing open vocabulary detectors leverage VLM backbones to enable generalisation to new object classes, embedded within Faster R-CNN-style architectures [11, 12, 15, 20, 22, 31, 41, 43, 50, 54, 55] or DETR-style architectures [44, 48]. CLIP [33] is frequently utilised by these open vocabulary detectors, directly included in the architecture via its text encoder [11, 12, 15, 22, 31, 41, 43, 48, 54, 55] or in its entirety [20, 44], or used as a supervision signal during training [11, 12, 15, 31, 41, 43, 48, 54]. We refer the reader to [42] for a comprehensive survey on open vocabulary learning.

2.3 Open-set Recognition

Scheirer et al. [37] introduced open-set recognition to challenge the prevailing closed-set assumptions of image recognition systems. *Closed-set* models only test on samples from a predefined set of classes that were known during model training [37]. In *open-set* recognition, samples can instead come from an expansive range of classes that cannot be predefined, and therefore are unknown during model training [37]. Open-set recognition tests the ability of models to identify and reject unknown samples that do not belong to one of the training classes.

The open-set concept has been explored widely in computer vision, including in image classification [4, 7, 28, 30, 40, 45, 53], object detection [10, 16, 26, 27, 56], image segmentation [6, 17, 32] and other vision tasks [3, 24, 25, 29, 35].

Our paper is the first to systematically investigate the open-set problem for VLMs, and our findings challenge the assumption that the open-set problem has been solved for VLMs due to their exposure to internet-scale training data. The following sections define the closed-set assumptions existing in current VLMs, demonstrate that VLMs continue to be vulnerable to open-set conditions and evaluate a number of baseline strategies for open-set recognition with VLMs.

3 Problem Definition

The core concept behind VLMs, mapping images and text into a joint embedding space, makes them powerful foundations for image classification and object detection. Classification is performed by comparing an image embedding \mathbf{e}^I with text embeddings $\mathcal{E}^q = \{\mathbf{e}_1^q, \dots, \mathbf{e}_N^q\}$ obtained from a set of query labels $\mathcal{Q} = \{q_1, \dots, q_N\}$ which are class names such as “dog”¹. The similarities between the image embedding and all text embeddings from \mathcal{Q} are measured by a similarity function $\mathbf{s}^q = \text{sim}(\mathbf{e}^I, \mathcal{E}^q)$, e.g. using the cosine similarity. The text embedding with the highest similarity to the image embedding, i.e. $\text{argmax} \mathbf{s}^q$, determines the predicted class label. The process is the same for object detection, with \mathbf{e}^I now representing the embedding of an image region corresponding to an object proposal.

Open-set and Closed-set: While the pre-VLM definition of the closed-set and open-set concepts focused on the challenges arising from a finite *training* set [37], our updated definition emphasises the limitations of the finite *query* set: The closed-set assumption in VLMs is that all classes seen during testing are predefined prior to testing and included in the query set. An open-set situation arises when the VLM encounters an object or image class that is not part of this predefined query set \mathcal{Q} .

3.1 Baseline Approaches to Open-Set Recognition for VLMs

In an open-set situation, the VLM should be able to identify an input not represented by the query set and reject it as **unknown**, instead of misclassifying it as one of the query classes from \mathcal{Q} . Similar to the pre-VLM open-set literature, there are multiple ways of recognising an input as unknown [13]. We investigate baseline methods using predictive uncertainty and different approaches of negative embeddings.

Using Uncertainty: Different measures of predictive uncertainty ψ can be used to reject inputs as **unknown**, for example based on the class similarities, e.g.

¹ Most methods insert an additional step here and generate a set of prompts from each query label q_i , such as {“A photo of a dog”, “A picture of a dog”, ...} and compare the embeddings of these prompts with \mathbf{e}^I . For clarity of notation, we omit this step.

$\psi^{\text{Cosine}} = \max(\mathbf{s}^q)$ or $\psi^{\text{Softmax}} = \max(\text{softmax}(\mathbf{s}^q))$, or the entropy of the softmax similarities $\psi^{\text{Entropy}} = -H(\text{softmax}(\mathbf{s}^q))$. Only inputs that pass a threshold test $\psi \geq \theta$ are classified according to $\text{argmax } \mathbf{s}^q$. Otherwise, the input is declared to be open-set and unknown.

Dedicated Negative Embeddings: Another approach is to create a set of augmented query embeddings $\tilde{\mathcal{E}}^q = \mathcal{E}^q \cup \mathcal{E}^n$ by adding M dedicated **negative** class embeddings $\mathcal{E}^n = \{\mathbf{e}_1^n, \dots, \mathbf{e}_M^n\}$. The input image or detection can then be rejected as **unknown** if the image embedding is most similar to one of those **negative** embeddings \mathbf{e}_i^n rather than a query class embedding \mathbf{e}_j^q . These negative embeddings can be created in one of two ways: (1) by creating **negative words**, which are then passed through the VLM text encoder to create the negative embeddings, or (2) by directly creating **negative embeddings**.

We investigate two baseline options for negative embeddings: a **random words** method, where we generate random strings with random length between 2-8 characters (e.g. “braxl”); and a **random embeddings** method, where we randomly draw M embedding vectors from a Gaussian distribution, i.e. $\mathbf{e}_i^n \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, using the mean and standard deviation from the query embedding vectors \mathcal{E}^q . We compare the effectiveness of these approaches, including for different numbers of M random negative queries, in our experiments. Interestingly, many open vocabulary object detectors already use a negative query approach, though typically with a only single embedding vector of all zeroes [22, 44, 50, 54, 55].

4 Evaluation Protocol

4.1 Creating an Open-set Recognition Dataset for VLMs

We now introduce how to use existing image classification or object detection datasets to evaluate open-set recognition. In short, we test all images twice: once in a standard closed-set manner with all dataset classes included in the query set (see Fig. 1 top-right), and once in an open-set manner with a query set containing all dataset classes not present in the image (see Fig. 1 bottom-right).

Classification and object detection datasets have a pre-defined set of K labelled target classes \mathcal{L} . In image classification, there is a single corresponding class label $y \in \mathcal{L}$ per image. In object detection, there are multiple ground truth objects in each image and each has a corresponding class label. For both tasks, we can define a set \mathcal{Y} that contains the ground truth class labels for each image: \mathcal{Y} contains a single label in classification and multiple labels in object detection.

Testing Closed-set Recognition: We first test images in a standard manner to establish the closed-set performance of the VLM and identify all *true positive (TP)* predictions. This involves testing all images with a query set identical to the dataset class list, $\mathcal{Q} = \mathcal{L}$. To identify the TP predictions, we follow standard protocol for classification and object detection datasets – in classification, a TP has a predicted class matching the ground-truth label; in object detection, this involves a object-detection assignment process that requires correct object

classification and localisation (see Supplementary Sec. S1). TPs represent correct closed-set predictions that should not be rejected as **unknown**, and we collect the uncertainty measures ψ associated with all TPs for later evaluation.

For object detection datasets, unlabelled objects are present in the background of the image [10]. These unlabelled objects are not in the query set \mathcal{Q} , and therefore meet our definition for open-set objects. Detections of these open-set objects will present as false positives (FPs) in a closed-set test, however not all FPs arise from open-set objects – FPs in object detection can also arise due to poor object localisation, misclassification between query classes, or duplicate detections of a query class (see [5]). Rather than attempting to categorise different FP types to distinguish open-set errors, our approach definitively identifies open-set errors for both object detection and image classification.

Testing Open-set Recognition: To evaluate open-set recognition, we test each image in the dataset with a modified query set, $\tilde{\mathcal{Q}}$ that contains all the class labels not in the image ground-truth class list, $\tilde{\mathcal{Q}} = \mathcal{L} - \mathcal{Y}$. This ensures the image only contains object classes that are not in the modified query set – every single prediction must be an *open-set error (OSE)*². In image classification, this modifies the query set to have $K - 1$ labels. In object detection, the modified query set size depends on the number of unique object classes in the image. We collect the uncertainty measures ψ of all OSE predictions for evaluation.

4.2 Metrics

We use performance metrics that measure the ability of the VLM to reject OSEs (with high uncertainty or with a **negative** class), while maintaining the closed-set task performance.

Precision-Recall Curves: Open-set recognition can be formulated as a binary classification task that uses uncertainty as the decision threshold. Correct closed-set predictions (introduced as TPs in Sec. 4.1) and open-set errors (OSEs) are considered as the positive and negative class respectively. We construct precision-recall curves, which are well-suited to this problem as they are affected by the size of the negative class (OSEs) [36]³ as well as prediction uncertainty. High precision indicates that little-to-none negatives (OSEs) remain after the uncertainty thresholding. High recall indicates that most positives (TPs) remain after the uncertainty thresholding. We summarise the performance of the precision-recall curves with 3 core metrics: (1) **Area under the PR curve (AuPR)**, (2) **Precision at 95% Recall**, and (3) **Recall at 95% Precision**.

Task Metrics: We use the established task metrics used to evaluate classification and detection – top-1 accuracy [34] and mean Average Precision (mAP) [23]

² This assumes accurate labelling of ground-truth classes, though most datasets will contain some label error. This approach is unsuitable for weakly-labelled datasets.

³ ROC curves do not have this property – we discuss this further in Sec. S3 of the Supplementary Material where we report AuROC for the object detection experiments.

respectively. We evaluate mAP at an IoU threshold of 0.5 – we are primarily interested in the classification ability of the detector (rather than localisation) and this is standard protocol in open vocabulary detection [15, 22, 44, 48, 50, 54, 55].

Absolute Counts: We also report the absolute counts of TPs and OSEs. This can be particularly useful for object detection, where increased task performance does not necessarily indicate increased TPs and detectors predict variable numbers of OSEs. We stress that *absolute counts should not be considered alone as a performance indicator*, e.g. an increased number of OSEs can be acceptable if all OSEs have very high uncertainty.

4.3 VLM Baselines

We test with six VLM classifiers – CLIP [33], ALIGN [19], CoCa [46], ImageBind [14], SigLIP [52] and LanguageBind [58] – and seven open vocabulary object detectors – OVR-CNN [50], ViLD [15], OV-DETR [48], RegionCLIP [54], Detic [55], VLDet [22] and CORA [44] (see Supplementary Sec. S2 for implementation details). These VLMs cover a range of architectures and training paradigms and have publicly-available implementations. Our goal with testing this range of VLMs is not to find the “best” VLM for open-set recognition, but instead to exhibit the general vulnerability of VLMs to open-set conditions. For each VLM, we extract the prediction *Softmax* score, *Cosine* similarity, and *Entropy* of the softmax distribution as baseline uncertainty measures. Some VLMs use Sigmoid activation rather than Softmax [22, 44, 48, 52, 55] – for these methods, we only report the *Sigmoid* score as its results are identical to the cosine similarity and the entropy is unsuitable as the scores are not normalised to a distribution. We do not impose a minimum score threshold on any predictions, as the AuPR metric captures performance over all possible thresholds.

4.4 Datasets

We focus on general object recognition and test with the large-scale and prevailing datasets used to benchmark image classification and object detection: For image classification, we test with the ImageNet1k validation dataset [9, 34]. It contains 50,000 images from 1000 classes and is publicly available for research. For object detection, we test with the MS COCO 2017 validation dataset [23]. It contains 4,952 images with 36,781 annotated objects from 80 classes and is publicly available for research. We show results for domain-specific classification datasets in the Supplementary Sec. S4.

5 Experiments and Results

5.1 State-of-the-art VLMs Perform Poorly in Open-set Conditions.

Classification: Tab. 1 shows that all six VLM classifiers perform poorly for open-set recognition. For scenarios that require high recall of true positive

Table 1: The open-set performance of state-of-the-art VLM classifiers when tested on ImageNet1k. Arrows indicate direction of optimal performance and - indicates operating point could not be achieved.

Classifier	Uncertainty	AuROC	AuPR	P@95R	R@95P	TP	OSE	Accuracy
		↑	↑	↑	↑	↑	↓	↑
CLIP [33]	Softmax	79.7	71.5	46.2	3.4			
	Cosine	72.2	61.3	42.5	0.1	31,023	50,000	62.1
	Entropy	80.2	72.3	45.4	2.8			
ALIGN [19]	Softmax	81.0	74.7	48.1	8.3			
	Cosine	72.4	62.1	44.8	-	32,618	50,000	65.2
	Entropy	80.9	75.0	46.8	8.6			
CoCa [46]	Softmax	79.0	71.7	46.0	3.4			
	Cosine	73.2	63.0	43.3	0.1	31,725	50,000	63.4
	Entropy	80.5	73.0	46.8	3.5			
ImageBind [14]	Softmax	82.8	79.1	52.8	5.0			
	Cosine	79.2	73.9	50.7	0.2	38,405	50,000	76.8
	Entropy	84.3	80.1	56.3	5.6			
SigLIP [52]	Sigmoid	81.0	76.2	52.1	2.2	37,851	50,000	75.7
LanguageBind [58]	Softmax	83.9	80.7	54.7	10.5			
	Cosine	81.6	77.4	52.8	0.8	39,243	50,000	78.5
	Entropy	85.4	81.7	58.4	11.1			

predictions (i.e. 95% recall), approximately every second prediction returned is an open-set error (i.e. precision between 46.2% and 58.4%). For scenarios that require high precision of predictions (i.e. 95% precision), less than 12% of the true positive predictions are retained (i.e. recall between 3.4% and 11.1%). Comparing across uncertainty types, the VLM classifiers show better uncertainty performance when using the predicted Softmax score or Entropy rather than the predicted Cosine similarity. This suggests raw similarity in VLM feature space is not the best indicator of open-set error, with measures of relative similarity comparative to other classes offering better performance.

Object Detection: All seven of the tested VLM object detectors are vulnerable to open-set errors – see Tab. 2. Unlike VLM classifiers, most VLM detectors already contain a negative query to capture objects not in the query set (often referred to as the *background* class). Despite this, the VLM detectors produce open-set error counts in the range of 100,000 to 1,500,000 when tested on only 4,952 images, i.e. between 20 and 300 open-set error per image on average.

High open-set error counts are not inherently a problem if they are produced with high uncertainty. Yet for all tested detectors, the baseline uncertainty methods cannot adequately distinguish true positive detections from open-set errors. When thresholding for high recall of true positives, *at best* only every fifth detection is a true positive (i.e. precision ranges between 5.9% and 21.9%). When

Table 2: The open-set performance of state-of-the-art open-vocabulary object detectors (with ResNet backbone indicated), tested on COCO. Arrows indicate the direction of optimal performance.

Detector	Uncertainty	Negative	AuPR	P@95R	R@95P	TP	OSE	mAP
		Embedding	↑	↑	↑	↑	↓	↑
OVR-CNN (R50) [50]	Softmax	Zero-Emb	75.4	13.0	42.9			
	Cosine	Zero-Emb	77.2	15.3	43.6	13,544	190,146	34.7
	Entropy	Zero-Emb	63.2	7.5	43.4			
ViLD (R152) [15]	Softmax	“background”	60.1	7.5	15.4			
	Cosine	“background”	33.1	6.9	0.1	16,011	1,485,600	46.4
	Entropy	“background”	62.2	8.4	16.3			
OV-DETR (R50) [48]	Sigmoid	None	75.8	5.9	47.3	15,818	1,485,600	43.0
RegionCLIP (R50) [54]	Softmax	Zero-Emb	73.1	9.9	39.4			
	Cosine	Zero-Emb	74.0	7.1	40.9	15,741	493,940	44.6
	Entropy	Zero-Emb	35.4	3.9	17.0			
Detic (R50) [55]	Sigmoid	Zero-Emb	72.6	8.1	42.2	14,670	495,200	39.1
VLDet (R50) [22]	Sigmoid	Zero-Emb	79.8	21.9	48.5	14,751	112,923	40.6
CORA (R50) [44]	Sigmoid	Zero-Emb	67.2	8.1	24.7	13,763	495,200	31.6

thresholding for high precision, *at best* less than half of true positives are recalled (i.e. recall ranges between 16.3% and 48.5%). We include some qualitative examples of highly confident open-set errors in the Supplementary Sec. S6.

Notably, the precision-recall characteristics for VLM classification and object detection differ significantly – VLM classifiers achieve better performance at high recall, whereas VLM detectors achieve better performance at high precision. This suggests that while the VLM classifiers struggle with over-confident open-set errors, the VLM detectors suffer from under-confident true positives. We include histograms that visualise this trend in the Supplementary Sec. S7.

We also consider how the performance of uncertainty differs for identifying open-set error versus closed-set misclassifications, with results and a detailed discussion in the Supplementary Sec. S5. We find that the performance of softmax uncertainty is lower for open-set error than for closed-set error in both classification and detection tasks. In classification, open-set performance is particularly decremented when thresholding for high precision – likely due to a greater proportion of overconfident open-set error – whereas detection primarily degrades with increased numbers of errors, with ViLD for example producing $11.7\times$ more open-set than closed-set misclassifications.

5.2 Can Negative Queries Improve Open-set Performance?

Classification: Negative class queries provide the VLM classifiers with the ability to select “unknown” when testing an input. When used in larger quantities, negative queries can capture and remove significant portions of open-set error, yet this happens in trade-off with a reduction in closed-set performance –

Table 3: The open-set performance of state-of-the-art VLM classifiers on ImageNet1k when used with negative random words (R-W) or random embeddings (R-E).

Classifier	Negative Embedding	AuPR ↑	P@95R ↑	R@95P ↑	TP ↑	OSE ↓	Accuracy ↑
CLIP [33] (Softmax)	None	71.5	46.2	3.4	31,023	50,000	62.1
	500 R-W	73.4	48.1	3.3	29,961	41,428	59.9
	2500 R-W	72.8	49.8	1.6	28,657	35,092	57.3
	500 R-E	71.9	47.2	2.5	30,160	44,673	60.3
	2500 R-E	71.7	47.1	0.9	29,291	40,497	58.6
ALIGN [19] (Softmax)	None	74.7	48.1	8.3	32,618	50,000	65.2
	500 R-W	75.3	48.9	10.4	31,731	43,198	63.5
	2500 R-W	74.6	49.5	8.9	30,588	37,738	61.2
	500 R-E	74.9	48.7	8.6	32,342	47,228	64.7
	2500 R-E	75.1	49.7	7.2	31,462	41,210	62.9
CoCa [46] (Softmax)	None	71.7	46.0	3.4	31,725	50,000	63.4
	500 R-W	75.4	49.6	0.1	30,276	39,528	60.6
	2500 R-W	75.9	50.9	6.0	29,107	34,360	58.2
	500 R-E	71.7	46.0	3.4	31,723	49,966	63.4
	2500 R-E	71.7	46.1	3.4	31,719	49,866	63.4
ImageBind [14] (Softmax)	None	79.1	52.8	5.0	38,405	50,000	76.8
	500 R-W	83.8	63.1	7.7	35,859	29,172	71.7
	2500 R-W	84.6	66.1	8.3	34,179	22,872	68.4
	500 R-E	81.0	57.3	9.0	36,966	38,636	73.9
	2500 R-E	81.7	59.0	11.2	35,954	33,349	71.9
SigLIP [52] (Sigmoid)	None	76.2	52.1	5.2	37,851	50,000	75.7
	500 R-W	77.6	56.7	2.3	36,111	33,287	72.2
	2500 R-W	78.8	60.2	2.4	34,507	26,110	69.0
	500 R-E	84.7	73.0	3.6	22,772	9,570	45.5
	2500 R-E	90.1	83.9	7.9	10,561	2,258	21.1
LanguageBind [58] (Softmax)	None	80.7	54.7	10.5	39,243	50,000	78.5
	500 R-W	84.7	64.9	13.1	36,596	28,337	73.2
	2500 R-W	85.4	68.2	11.7	34,540	21,230	69.1
	500 R-E	80.8	54.9	10.8	39,226	49,646	78.5
	2500 R-E	80.9	55.1	10.9	39,210	49,032	78.4

see Tab. 3. For example, ImageBind using 2,500 random word negative queries effectively halves the open-set error, but this occurs in trade-off with a closed-set accuracy reduction of 8.4% (approximately 4,000 of the closed-set images are incorrectly classified as one of the random words).

While negative queries noticeably reduce the raw open-set error count, this is not always mirrored with increased performance in the open-set uncertainty metrics. We present the change in closed-set and open-set performance against the number of negative query points in Fig. 2. Generally the use of negative random words appears the most effective, with CoCa, ImageBind, SigLIP and LanguageBind showing improved open-set performance with minimal loss of closed-set performance. For some of the classifiers, there is negligible change in open-set performance despite the reduction in open-set error. We explore this further in the Supplementary Sec. S8, where we show that in some VLMs the negative

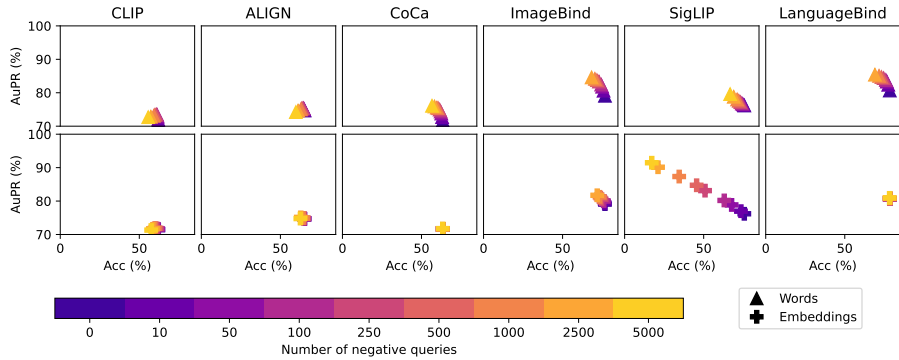


Fig. 2: The trade-off between closed-set and open-set performance with increasing number of negative queries for different VLM classifiers.

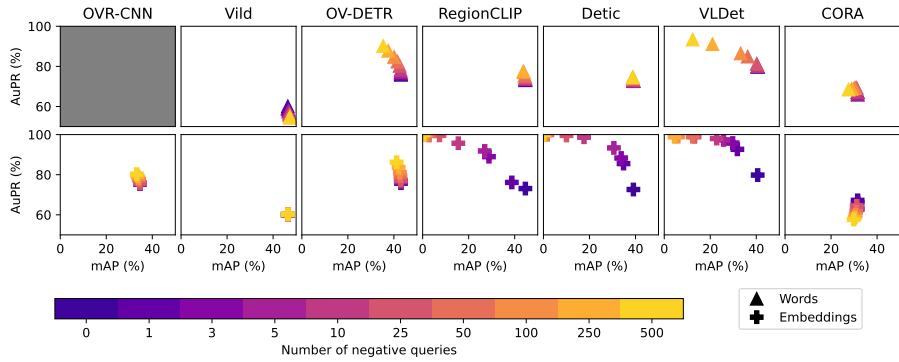


Fig. 3: The trade-off between closed-set and open-set performance with increasing number of negative queries for different VLM object detectors.

queries only capture and remove “easy” open-set errors, i.e. open-set errors that already have high uncertainty.

Object Detection: The VLM detectors do not show a consistent response to the use of negative queries. Different detectors show different sensitivities to negative query counts, different trade-off rates between closed-set and open-set performance, and differences in effectiveness of embedding versus word negative query types – see Tab. 4 and Fig. 3. Looking at Fig. 3, we can consider the sensitivity, trade-off rate and response of the detectors to negative queries: greater performance changes with increasing queries indicates sensitivity; large decreases in mAP with more queries indicates increased trade-off rates between open-set and closed-set performance; and a positive response to the negative queries presents as an increase in AuPR when increasing queries (negative response decreases AuPR).

Table 4: The open-set performance of state-of-the-art open vocabulary object detectors (with ResNet backbone indicated) when additional negative embeddings are included, tested on COCO. Arrows indicate the direction of optimal performance. Negative embeddings introduced as words are highlighted gray.

Detector	Negative Embedding	AuPR ↑	P@95R ↑	R@95P ↑	TP ↑	OSE ↓	mAP ↑
OVR-CNN (R50) [50]	Zero-Emb	75.4	13.0	42.9	13,544	190,146	34.7
	+5 Rand-Embs	75.6	13.1	43.1	13,543	186,120	34.7
	+100 Rand-Embs	78.3	18.2	45.6	13,129	109,267	34.0
ViLD (R152) [15]	“background”	60.1	7.5	15.4	16,011	1,485,600	46.4
	+5 Rand-Words	57.5	6.0	13.4	15,926	1,400,612	46.6
	+100 Rand-Words	55.0	2.7	10.3	15,772	1,088,725	47.3
	+5 Rand-Embs	60.1	7.5	15.2	16,011	1,485,231	46.4
	+100 Rand-Embs	60.2	7.6	15.2	16,019	1,478,806	46.4
OV-DETR (R50) [48]	None	75.8	5.9	47.3	15,818	1,485,600	43.0
	+5 Rand-Words	77.2	8.1	48.3	15,510	923,398	43.0
	+100 Rand-Words	84.7	27.0	55.7	13,433	239,118	40.1
	+5 Rand-Embs	76.6	6.6	47.8	15,652	1,178,827	43.0
	+100 Rand-Embs	81.0	18.2	50.9	14,699	356,380	42.6
RegionCLIP (R50) [54]	Zero-Emb	73.1	9.9	39.3	15,752	493,949	44.6
	+5 Rand-Words	73.8	11.1	39.7	15,662	322,596	44.6
	+100 Rand-Words	77.0	17.5	43.1	15,178	154,262	44.0
	+5 Rand-Embs	91.8	77.2	44.6	8,828	5,062	27.0
	+100 Rand-Embs	99.9	99.3	83.6	287	2	1.6
Detic (R50) [55]	Zero-Emb	72.6	8.1	42.2	14,670	495,200	39.1
	+5 Rand-Words	72.8	8.5	42.4	14,637	451,633	39.1
	+100 Rand-Words	74.5	11.5	43.5	14,287	265,373	38.8
	+5 Rand-Embs	93.4	74.5	61.8	10,070	6,346	30.6
	+100 Rand-Embs	100.0	100.0	100.0	371	0	0.7
VLDet (R50) [22]	Zero-Emb	79.8	21.9	48.4	14,751	112,954	40.6
	+5 Rand-Words	80.1	22.3	48.7	14,672	111,893	40.5
	+100 Rand-Words	86.5	36.7	58.4	11,132	50,122	33.2
	+5 Rand-Embs	97.3	91.3	84.9	8,464	1,382	25.8
	+100 Rand-Embs	99.0	99.3	3.7	2,531	22	5.9
CORA (R50) [44]	Zero-Emb	67.2	8.1	24.9	13,745	495,200	31.5
	+5 Rand-Words	66.3	7.6	27.0	13,700	413,636	31.6
	+100 Rand-Words	69.1	10.1	28.6	13,722	329,393	30.1
	+5 Rand-Embs	63.1	6.2	21.8	13,689	495,200	31.5
	+100 Rand-Embs	59.9	6.0	16.7	13,680	495,174	30.6

OV-DETR is the only detector to show a clear positive response, low trade-off rate, and high sensitivity to the negative query approach, particularly with negative embeddings. While OVR-CNN (embeddings), RegionCLIP (words), Detic (words) also show a positive response and low trade-off rate, this is with very little sensitivity to increasing numbers of negative queries. In contrast, the open-set performance of Detic, RegionCLIP, and VLDet are very sensitive to negative

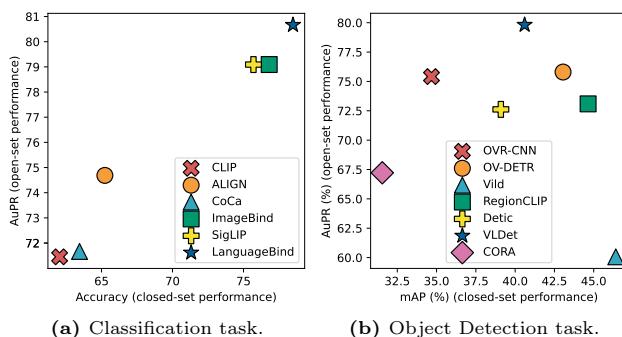


Fig. 4: The relationship between closed-set and open-set performance in VLMs.

embeddings, but at a rapid trade-off with closed-set performance. Both ViLD and CORA show a negative response or no sensitivity to negative queries.

It is tempting to relate these trends to the architecture or training paradigm of the VLM detectors, however we could not identify many common characteristics. RegionCLIP, Detic and VLDet all learn from auxiliary datasets that have been pseudo-labelled with regions [22, 54, 55], although this does not clearly indicate a reason for high sensitivity and trade-off rate for negative embeddings. ViLD [15] and CORA [44] both heavily rely on CLIP’s visual feature encoding (either by using CLIP directly [44] or knowledge distillation [15]) – their poor response is consistent with CLIP in the classification experiments on negative queries. Yet OV-DETR also uses CLIP’s visual features during training to condition the object proposals [48], but exhibits the best response to negative queries.

5.3 Is a good classifier all you need for VLM Open-set Recognition?

In traditional open-set recognition, Vaze et. al [40] identified a strong positive correlation between closed-set and open-set performance. Testing this theory for VLM open-set recognition, we compare classifier and object detector closed-set and open-set performance in Fig. 4.

Classification: The correlation observed by [40] appears to hold for VLM classifiers, with greater accuracy (closed-set performance) relating to greater AuPR scores (open-set performance). It is worth noting that AuPR is a summary metric and this behaviour may not hold when examining performance at different operating points on the precision-recall curves.

Object Detection: There is not a clear correlation between a VLM object detector’s closed-set and open-set performance. ViLD in particular challenges this correlation, with the greatest mAP yet worst AuPR of all the tested detectors.

5.4 How Does Query Set Size Impact Performance?

In Fig. 5, we test performance of the VLM classifiers on ImageNet while progressively increasing the query set size. We select the additional classes from the

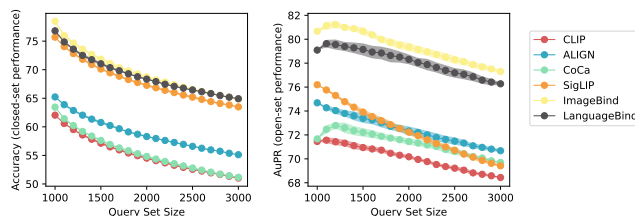


Fig. 5: Influence of the query set size on closed-set and open-set performance on ImageNet1k classification, showing mean and standard deviation from 10 random seeds.

WordNet hierarchy by choosing leaf nodes that are related, but not too closely related, to existing ImageNet classes (i.e. share a grandparent in the hierarchy but do not share a direct parent). Increasing the number of target classes and thus the query set size decrements both the closed-set and open-set performance of a VLM classifier. This indicates that testing VLMs with an entire vocabulary of nouns may not be a feasible approach to avoid open-set recognition.

6 Discussion

Open Challenges: The tested baseline uncertainty measures were ineffective at identifying open-set errors, indicating a need to research better alternatives. While uncertainty representations for VLMs have been proposed [18, 39], their utility for open-set recognition remains unexplored.

While the use of negative random embeddings was able to reduce open-set error, large numbers were typically required to have any effect and this was often as the cost of falsely captured true positive predictions. Interestingly, some negative embeddings were much more effective at capturing open-set errors than others (e.g. the random string “awy” tested with ImageBind, see the Supplementary Sec. S9). Identifying the characteristics of these embeddings, learning them through prompt-tuning [8, 11, 12, 21, 57] or introducing auto-labelling techniques [49] are promising directions.

Conclusions: VLMs are increasingly applied in diverse real-world applications, from robotics to medical image analysis. We hope that by highlighting their weaknesses, we encourage further research to improve these models and enhance their positive societal impact. We have clearly demonstrated that VLMs for open-vocabulary image classification and object detection introduce closed-set assumptions by relying on a finite query set. We thus refuted the assumption that vision-language models (VLMs) are inherently open-set models due to their training on extensive internet-scale datasets. While the investigated baseline approaches to identify open-set errors showed promise, we found none of them sufficiently robust and reliable for safety-critical applications. Our new benchmark and evaluation protocol is designed to foster and support this research. Understanding VLMs’ limitations ultimately reduces risks, improves transparency, and ensures safer deployment in critical applications.

References

1. Alayrac, J.B., Donahue, J., Luc, P., Miech, A., Barr, I., Hasson, Y., Lenc, K., Mensch, A., Millican, K., Reynolds, M., et al.: Flamingo: a visual language model for few-shot learning. *Advances in Neural Information Processing Systems* **35**, 23716–23736 (2022)
2. Amodei, D., Olah, C., Steinhardt, J., Christiano, P., Schulman, J., Mané, D.: Concrete problems in ai safety. *arXiv preprint arXiv:1606.06565* (2016)
3. Bao, W., Yu, Q., Kong, Y.: Evidential deep learning for open set action recognition. In: *Proceedings of the IEEE/CVF International Conference on Computer Vision*. pp. 13349–13358 (2021)
4. Bendale, A., Boulton, T.E.: Towards open set deep networks. In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. pp. 1563–1572 (2016)
5. Bolya, D., Foley, S., Hays, J., Hoffman, J.: Tide: A general toolbox for identifying object detection errors. In: *Computer Vision—ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part III 16*. pp. 558–573. Springer (2020)
6. Cen, J., Yun, P., Cai, J., Wang, M.Y., Liu, M.: Deep metric learning for open world semantic segmentation. In: *Proceedings of the IEEE/CVF International Conference on Computer Vision*. pp. 15333–15342 (2021)
7. Chen, G., Peng, P., Wang, X., Tian, Y.: Adversarial reciprocal points learning for open set recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence* **44**(11), 8065–8081 (2021)
8. Cho, E., Kim, J., Kim, H.J.: Distribution-aware prompt tuning for vision-language models. In: *Proceedings of the IEEE/CVF International Conference on Computer Vision*. pp. 22004–22013 (2023)
9. Deng, J., Dong, W., Socher, R., Li, L.J., Li, K., Fei-Fei, L.: Imagenet: A large-scale hierarchical image database. In: *2009 IEEE conference on computer vision and pattern recognition*. pp. 248–255. Ieee (2009)
10. Dhamija, A., Gunther, M., Ventura, J., Boulton, T.: The overlooked elephant of object detection: Open set. In: *The IEEE Winter Conference on Applications of Computer Vision*. pp. 1021–1030 (2020)
11. Du, Y., Wei, F., Zhang, Z., Shi, M., Gao, Y., Li, G.: Learning to prompt for open-vocabulary object detection with vision-language model. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. pp. 14084–14093 (2022)
12. Feng, C., Zhong, Y., Jie, Z., Chu, X., Ren, H., Wei, X., Xie, W., Ma, L.: Prompt-det: Towards open-vocabulary detection using uncurated images. In: *European Conference on Computer Vision*. pp. 701–717. Springer (2022)
13. Geng, C., Huang, S.j., Chen, S.: Recent advances in open set recognition: A survey. *IEEE transactions on pattern analysis and machine intelligence* **43**(10), 3614–3631 (2020)
14. Girdhar, R., El-Nouby, A., Liu, Z., Singh, M., Alwala, K.V., Joulin, A., Misra, I.: Imagebind: One embedding space to bind them all. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. pp. 15180–15190 (2023)
15. Gu, X., Lin, T.Y., Kuo, W., Cui, Y.: Open-vocabulary object detection via vision and language knowledge distillation. In: *International Conference on Learning Representations* (2022)

16. Han, J., Ren, Y., Ding, J., Pan, X., Yan, K., Xia, G.S.: Expanding low-density latent regions for open-set object detection. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 9591–9600 (2022)
17. Hwang, J., Oh, S.W., Lee, J.Y., Han, B.: Exemplar-based open-set panoptic segmentation network. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 1175–1184 (2021)
18. Ji, Y., Wang, J., Gong, Y., Zhang, L., Zhu, Y., Wang, H., Zhang, J., Sakai, T., Yang, Y.: Map: Multimodal uncertainty-aware vision-language pre-training model. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 23262–23271 (2023)
19. Jia, C., Yang, Y., Xia, Y., Chen, Y.T., Parekh, Z., Pham, H., Le, Q., Sung, Y.H., Li, Z., Duerig, T.: Scaling up visual and vision-language representation learning with noisy text supervision. In: Meila, M., Zhang, T. (eds.) Proceedings of the 38th International Conference on Machine Learning. Proceedings of Machine Learning Research, vol. 139, pp. 4904–4916. PMLR (18–24 Jul 2021), <https://proceedings.mlr.press/v139/jia21b.html>
20. Kuo, W., Cui, Y., Gu, X., Piergiovanni, A., Angelova, A.: F-vlm: Open-vocabulary object detection upon frozen vision and language models. In: International Conference on Learning Representations (2023)
21. Li, H., Zhang, R., Yao, H., Song, X., Hao, Y., Zhao, Y., Li, L., Chen, Y.: Learning domain-aware detection head with prompt tuning. *Advances in Neural Information Processing Systems* **36** (2024)
22. Lin, C., Sun, P., Jiang, Y., Luo, P., Qu, L., Haffari, G., Yuan, Z., Cai, J.: Learning object-language alignments for open-vocabulary object detection. In: International Conference on Learning Representations (2023)
23. Lin, T.Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., Zitnick, C.L.: Microsoft coco: Common objects in context. In: Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6–12, 2014, Proceedings, Part V 13. pp. 740–755. Springer (2014)
24. Liu, Y.C., Ma, C.Y., Dai, X., Tian, J., Vajda, P., He, Z., Kira, Z.: Open-set semi-supervised object detection. In: European Conference on Computer Vision. pp. 143–159. Springer (2022)
25. Maalouf, A., Jadhav, N., Jatavallabhula, K.M., Chahine, M., Vogt, D.M., Wood, R.J., Torralba, A., Rus, D.: Follow anything: Open-set detection, tracking, and following in real-time. *IEEE Robotics and Automation Letters* **9**(4), 3283–3290 (2024)
26. Miller, D., Nicholson, L., Dayoub, F., Sünderhauf, N.: Dropout sampling for robust object detection in open-set conditions. In: 2018 IEEE International Conference on Robotics and Automation (ICRA). pp. 3243–3249. IEEE (2018)
27. Miller, D., Sünderhauf, N., Milford, M., Dayoub, F.: Uncertainty for identifying open-set errors in visual object detection. *IEEE Robotics and Automation Letters* **7**(1), 215–222 (2022). <https://doi.org/10.1109/LRA.2021.3123374>
28. Oza, P., Patel, V.M.: C2ae: Class conditioned auto-encoder for open-set recognition. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 2307–2316 (2019)
29. Panareda Busto, P., Gall, J.: Open set domain adaptation. In: Proceedings of the IEEE international conference on computer vision. pp. 754–763 (2017)
30. Perera, P., Morariu, V.I., Jain, R., Manjunatha, V., Wigington, C., Ordonez, V., Patel, V.M.: Generative-discriminative feature representations for open-set recognition. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 11814–11823 (2020)

31. Pham, C., Vu, T., Nguyen, K.: Lp-ovod: Open-vocabulary object detection by linear probing. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision. pp. 779–788 (2024)
32. Pham, T., Do, T.T., Carneiro, G., Reid, I., et al.: Bayesian semantic instance segmentation in open set world. In: Proceedings of the European Conference on Computer Vision (ECCV). pp. 3–18 (2018)
33. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., Krueger, G., Sutskever, I.: Learning transferable visual models from natural language supervision. In: Meila, M., Zhang, T. (eds.) Proceedings of the 38th International Conference on Machine Learning. Proceedings of Machine Learning Research, vol. 139, pp. 8748–8763. PMLR (18–24 Jul 2021), <https://proceedings.mlr.press/v139/radford21a.html>
34. Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., Huang, Z., Karpathy, A., Khosla, A., Bernstein, M., et al.: Imagenet large scale visual recognition challenge. *International journal of computer vision* **115**, 211–252 (2015)
35. Saito, K., Yamamoto, S., Ushiku, Y., Harada, T.: Open set domain adaptation by backpropagation. In: Proceedings of the European conference on computer vision (ECCV). pp. 153–168 (2018)
36. Saito, T., Rehmsmeier, M.: The precision-recall plot is more informative than the roc plot when evaluating binary classifiers on imbalanced datasets. *PloS one* **10**(3), e0118432 (2015)
37. Scheirer, W.J., de Rezende Rocha, A., Sapkota, A., Boulton, T.E.: Toward open set recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence* **35**(7), 1757–1772 (2013). <https://doi.org/10.1109/TPAMI.2012.256>
38. Sünderhauf, N., Brock, O., Scheirer, W., Hadsell, R., Fox, D., Leitner, J., Uproft, B., Abbeel, P., Burgard, W., Milford, M., et al.: The limits and potentials of deep learning for robotics. *The International Journal of Robotics Research* **37**(4-5), 405–420 (2018)
39. Upadhyay, U., Karthik, S., Mancini, M., Akata, Z.: Problm: Probabilistic adapter for frozen vision-language models. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 1899–1910 (2023)
40. Vaze, S., Han, K., Vedaldi, A., Zisserman, A.: Open-set recognition: A good closed-set classifier is all you need? In: International Conference on Learning Representations (ICLR) (2022)
41. Wang, J., Zhang, H., Hong, H., Jin, X., He, Y., Xue, H., Zhao, Z.: Open-vocabulary object detection with an open corpus. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 6759–6769 (2023)
42. Wu, J., Li, X., Xu, S., Yuan, H., Ding, H., Yang, Y., Li, X., Zhang, J., Tong, Y., Jiang, X., Ghanem, B., Tao, D.: Towards Open Vocabulary Learning: A Survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence* (2024)
43. Wu, S., Zhang, W., Jin, S., Liu, W., Loy, C.C.: Aligning bag of regions for open-vocabulary object detection. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 15254–15264 (2023)
44. Wu, X., Zhu, F., Zhao, R., Li, H.: Cora: Adapting clip for open-vocabulary detection with region prompting and anchor pre-matching. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 7031–7040 (2023)
45. Yoshihashi, R., Shao, W., Kawakami, R., You, S., Iida, M., Naemura, T.: Classification-reconstruction learning for open-set recognition. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 4016–4025 (2019)

46. Yu, J., Wang, Z., Vasudevan, V., Yeung, L., Seyedhosseini, M., Wu, Y.: Coca: Contrastive captioners are image-text foundation models. *Transactions on Machine Learning Research* (2022), <https://openreview.net/forum?id=Ee277P3AYC>
47. Yuan, L., Chen, D., Chen, Y.L., Codella, N., Dai, X., Gao, J., Hu, H., Huang, X., Li, B., Li, C., et al.: Florence: A new foundation model for computer vision. *arXiv preprint arXiv:2111.11432* (2021)
48. Zang, Y., Li, W., Zhou, K., Huang, C., Loy, C.C.: Open-vocabulary detr with conditional matching. In: *European Conference on Computer Vision*. pp. 106–122. Springer (2022)
49. Zara, G., Roy, S., Rota, P., Ricci, E.: Autolabel: Clip-based framework for open-set video domain adaptation. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. pp. 11504–11513 (2023)
50. Zareian, A., Rosa, K.D., Hu, D.H., Chang, S.F.: Open-vocabulary object detection using captions. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. pp. 14393–14402 (2021)
51. Zhai, X., Wang, X., Mustafa, B., Steiner, A., Keysers, D., Kolesnikov, A., Beyer, L.: Lit: Zero-shot transfer with locked-image text tuning. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. pp. 18123–18133 (2022)
52. Zhai, X., et al.: Sigmoid loss for language image pre-training. In: *Intl. Conf. on Computer Vision*. pp. 11975–11986 (2023)
53. Zhang, H., Li, A., Guo, J., Guo, Y.: Hybrid models for open set recognition. In: *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part III 16*. pp. 102–117. Springer (2020)
54. Zhong, Y., Yang, J., Zhang, P., Li, C., Codella, N., Li, L.H., Zhou, L., Dai, X., Yuan, L., Li, Y., et al.: Regionclip: Region-based language-image pretraining. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. pp. 16793–16803 (2022)
55. Zhou, X., Girdhar, R., Joulin, A., Krähenbühl, P., Misra, I.: Detecting twenty-thousand classes using image-level supervision. In: *European Conference on Computer Vision*. pp. 350–368. Springer (2022)
56. Zhou, Z., Yang, Y., Wang, Y., Xiong, R.: Open-set object detection using classification-free object proposal and instance-level contrastive learning. *IEEE Robotics and Automation Letters* **8**(3), 1691–1698 (2023)
57. Zhu, B., Niu, Y., Han, Y., Wu, Y., Zhang, H.: Prompt-aligned gradient for prompt tuning. In: *Proceedings of the IEEE/CVF International Conference on Computer Vision*. pp. 15659–15669 (2023)
58. Zhu, B., Lin, B., Ning, M., Yan, Y., Cui, J., HongFa, W., Pang, Y., Jiang, W., Zhang, J., Li, Z., et al.: Languagebind: Extending video-language pretraining to n-modality by language-based semantic alignment. In: *Intl. Conf. on Learning Representations* (2023)